BIT-INTERLEAVED CODED MODULATION WITH ITERATIVE DEMAPPING AND DECODING FOR NON-COHERENT MIMO COMMUNICATION

BIT-INTERLEAVED CODED MODULATION WITH ITERATIVE DEMAPPING AND DECODING FOR NON-COHERENT MIMO COMMUNICATION

By

MOHAMED A. EL-AZIZY, B.Eng.Mgt.

AUGUST 2006

A Thesis

Submitted to the Department of Electrical & Computer Engineering

and the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree of

Master of Applied Science

McMaster University

MASTER OF APPLIED SCIENCE (2006) McMaster University
(Electrical & Computer Engineering) Hamilton, Ontario

TITLE: Bit-Interleaved Coded Modulation with Iterative Demapping and Decoding for Non-Coherent MIMO Communication

AUTHOR: Mohamed A. El-Azizy
B. Eng. Mgt. Electrical Engineering and Management
McMaster University, Hamilton, ON, Canada

SUPERVISOR: Timothy N. Davidson, Associate Professor

NUMBER OF PAGES: xii, 86

Dedications

To the memory of my grandfather,

to my grandmother,

my parents,

and my siblings.

Abstract

The goal of this thesis is the development of a computationally-efficient coded system that enables communication over the non-coherent Multiple-Input Multiple-Output (MIMO) flat-fading wireless channel at high data rates. The proposed signalling technique applies the principles of Bit-Interleaved Coded Modulation (BICM) with Iterative Demapping and Decoding (IDD) to non-coherent MIMO communication systems.

The principle of BICM is applied to a constellation that mimics the non-coherent capacity achieving distribution at high signal to noise ratios. The capacity achieving distribution is in the form of isotropically distributed unitary matrices, and the constellation can be represented by points on a Grassmannian manifold. A mapping technique that exploits the Grassmannian geometry is proposed. This mapping technique is based on the partitioning of the constellation into two subsets. The Grassmannian geometry also gives rise to an efficient list-based demapping algorithm that substantially reduces the computational complexity of the receiver while incurring some degradation in performance. For example, at a bit error rate (BER) of 10^{-4} the signal to noise ratio (SNR) performance degradation with respect to full constellation demapping is approximately 1.75 dB. A technique by which the decoder can augment the demapping list is proposed, and it is shown that the performance degradation of the efficient algorithm can be rendered insignificant (approximately 0.2 dB at a BER of 10^{-4}).

Finally, the performance of the proposed BICM-IDD using the Grassmannian constellation will be compared to that of a corresponding training-based BICM-IDD scheme. These simulations show that the proposed scheme can provide better performance at high data rates; e.g., for a data rate of 5/3 bits per channel use, the performance gap is almost 1 dB at BER of 10^{-4}.

Acknowledgments

Throughout my Master's program at McMaster University, I identify the input of several people whose contribution to my progress and life is evident.

I am particularly indebted to my father for his clear vision that has always guided me throughout my life. I truly appreciate his continuous support in all those years.

Without her love and belief in me, I would not have made it. She has always encouraged me through my struggles. Mom, I am so grateful to you.

As to my grandfather being an engineer, he taught me how to explore my options and be responsible for my decisions. Grandpa, I truly miss you and I wish you were still here to witness my graduation.

I am also grateful to my grandmother who taught me to be independent and to pursue my dreams. Her persistence and confidence in me have been a great source of encouragement.

As to my siblings, Rana and Omar, they have always covered up for my mistakes and were always there to cheer me up whenever I felt down.

I am also indebted to my friends H. Youssef, N. El-Nagar and M. Negm-Eldin for their continuous support.

I would like to thank the ECE department at McMaster University for the research facilities and the friendly environment. In particular, I would like to thank Mrs. C. Gies and Mrs. H. Jachna who were always a source of help and humor. I am also thankful to Mr. T. Greenlay and Mr. C. Coroiu for always being there everytime the

computers gave us trouble.

I am grateful to the Ontario Student Assistance Program (OSAP) for their financial support through the Ontario Graduate Scholarship (OGS).

I would like to thank all my colleagues at McMaster University who have helped to provide a supportive and exciting research atmosphere. I am especially indebted to my colleague and my friend, Dr. R. Gohary. Our discussions have contributed a great deal to my understanding. As a friend, he has always supported me in the time of need.

Finally, I am indebted to my supervisor Dr. T. N. Davidson for his guidance through the Master's program. His insightful comments and questions have enriched my knowledge and background in communication engineering. Dr. Davidson, your painstaking efforts have contributed greatly to the quality of the presentation of this work. It has been a great pleasure being your student and thank you.

Contents

List of Figures

Chapter 1

Introduction

The focus of this thesis is the development of a computationally-efficient coded scheme for communicating over non-coherent Multiple-Input Multiple-Output (MIMO) flat-fading wireless channel at rates close to the high-SNR ergodic capacity. Such schemes possess many desirable features for wireless systems that require reliable communication at high data rates. The proposed scheme is based on applying the principles of Bit-Interleaved Coded Modulation (BICM) with Iterative Demapping and Decoding (IDD) to non-coherent MIMO communication systems. The performance of the proposed system will then be compared to a corresponding training-based system. In this chapter, we will discuss the fundamentals of wireless communications, the principles of employing multiple antennas, the desirable properties of the non-coherent MIMO framework, and we will present the thesis outline.

1.1 Wireless Communication

In many practical applications maintaining a wireline communication can be both inconvenient and uneconomic [1]. For example, when communication sites are needed for a short period of time or when the terrain is too rugged to establish a wireline

connection. In some of these applications, wireless systems can offer substantial advantages [1]. In particular, they offer the potential of maintaining a communication link while both the transmitter and the receiver are mobile. In addition, wireless systems are not reliant on wireline infrastructure which can be expensive to construct and maintain [1].

Although the wireless medium is attractive, there are several factors that have to be carefully accounted for in the design of a wireless communication system. In a wireless medium, the transmitted signal is scattered off objects in its propagation paths resulting in multiple reflected versions of the signal at the receiver. The reflected signals travel along different propagation paths, and hence they experience different attenuation and time delays. Furthermore, the attenuation and delay on each path will usually vary with time, due to the relative motion of the transmitter, receiver and reflectors. This time variation is often called fading [2], although that description is perhaps most accurate when the time delays of each path are approximately the same. In that case, the amplitude of the received signal fluctuates, or "fades", depending on whether the components on each path add constructively or destructively.

Depending on the properties of both the channel and the transmitted signal, fading can be categorized into different classes. Fading can be categorized as being either slow or fast depending on the relation between the symbol duration and the coherence time of the channel. The coherence time of a channel is the time duration during which the channel remains essentially unchanged. If the coherence time is less than the symbol duration, then the signal is said to undergo fast fading. Otherwise, the transmitted signal is said to undergo slow fading [2]. Another attribute of fading depends on the relation between the bandwidth occupied by the signal and the coherence bandwidth of the channel, and this results in the categories of frequency-flat and frequency-selective fading. The coherence bandwidth of a channel is the frequency

interval over which the channel can be considered essentially constant. If the bandwidth occupied by the signal is much less than the coherence bandwidth of the channel then the signal is said to undergo frequency-flat fading. Otherwise, the transmitted signal undergoes frequency-selective fading [2]. Throughout this thesis, we will focus on the subclass of frequency flat slowly fading channels in which the channel remains constant for a period less than the coherence time, and then takes on independent values. This class of channels is referred to as the class of block-fading channels. This class is appropriate for modelling communication systems that employ time-division multiple access (TDMA) [3]. In these systems, each user is assigned a specific time slot in a signalling frame to transmit a signal. If the time interval spanned by each frame is significantly longer than the coherence time, then each block of bits of a certain user would experience an almost independent channel realization. Another communication system that can be represented by the block-fading model is one that employs frequency hopping. In such systems, the carrier frequency takes on a different value that is picked according to a pseudo-random sequence; e.g., Gold or Kasami sequences [2]. If the hopping bandwidth is significantly larger than the coherence bandwidth, the frequency response of the channel around each carrier frequency is approximately independent.

A technique to mitigate fading in wireless systems is to ensure that multiple representations of the same signal are transmitted over weakly correlated channel realizations and jointly processed at the receiver [4]. One way to construct such a scheme is by employing multiple antennas at the transmitter and/or receiver. In the case of weakly correlated channels, when a propagation path experiences deep fading in which the components of the signal propagating in different paths add almost destructively, the receiver can exploit the other approximately independent paths to improve the performance of the detection process. On the other hand, if the paths are highly correlated, when deep fading occurs in one path, it is likely that other

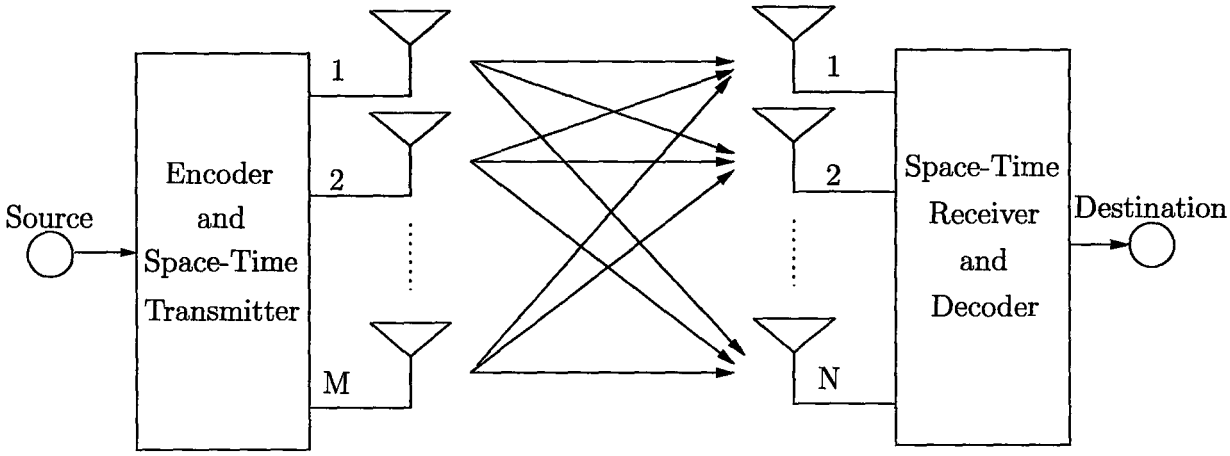


Figure 1.1: A generic multiple antenna system.

paths will be affected in a similar way leading to difficulties in the detection process. Fortunately, in environments with a high density of scatterers, the propagation paths are often highly uncorrelated [1]. Ideally, if there are many scatterers then the signal can be assumed to undergo independent fading along each propagation path [1]. Such a channel is referred as being "richly scattered". This can be the case when both the transmitter and receiver are located in an urban area and there is no direct line of sight. In order to ensure that the signal propagation paths are approximately independent in a richly scattered environment, field measurements show that the antennas at each end have to be separated by a distance at least equal to 0.5λ, where λ is the carrier wavelength [1]; see Figure 1.1. For example, if the carrier frequency is 6 GHz, the antennas should be separated by a distance greater than 2.5 cm.

In addition to fading, the received signals are also subject to thermal noise that is inherently generated in the electronic devices in the receiver and possibly interference from other users operating in the same frequency band [2]. However, the focus of this thesis is on single user communication systems. In the next section, we will focus on the principles of employing multiple antenna systems.

1.2 Multiple Antenna Wireless Systems

Unlike the single antenna model, in which the transmission is allocated in time slots, when communicating over a MIMO communication system the transmission is allocated in space and time slots. This allocation can be viewed as a two dimensional slot structure in which the space slot represents the antenna from which the data is transmitted and the time slot represents the time instants during which the transmission occurs.

In order to represent the space-time allocation of the transmitted signal, we will describe the basic building blocks of a standard MIMO communication system. A block diagram of this system is shown in Figure 1.2. In this standard system, the direct space-time encoding block of the generic MIMO system in Figure 1.1 is partitioned into a scalar encoder and a space-time mapper, and the direct space-time decoder is partitioned into a space-time demapper and a scalar decoder. Although this partitioning results in a loss in performance, it substantially simplifies the design and implementation of the both the transmitter and the receiver, while still offering the potential for reliable communication at high data rates [4–6]. The standard system in Figure 1.2 operates in the following way. The information source generates a stream of data bits that are passed into an encoder. The encoded bits are then fed to a MIMO space-time mapper in which the bits are divided into blocks and each block is mapped onto a space-time constellation matrix S. For example, one way to generate the space-time matrix S, is by further dividing the block of bits into sub-blocks and mapping each sub-block of bits onto a symbol s chosen from a standard constellation; e.g., Quadrature amplitude modulation (QAM) or phase shift keying (PSK). Then each mapped symbol s is allocated to a space-time slot to be transmitted on a certain antenna at a certain time; generating a space-time matrix S. At the destination, the receiver first detects the transmitted signal using a space-time demapper in which

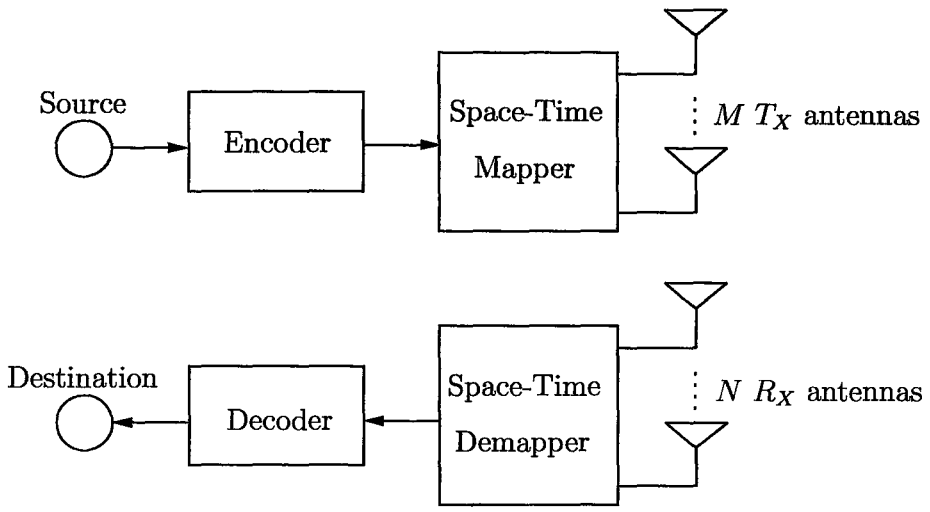


Figure 1.2: A standard MIMO communication system.

the received signal is demapped to produce a stream of bits (or "soft" information about these bits). This stream of bits is then passed to the decoder, which attempts to reproduce the original stream of data bits.

As previously mentioned, multiple antennas can be exploited to improve the reliability of the communication system. A simple way in which this can be accomplished is by transmitting a signal based on the same message from each antenna. The transmitted signals follow different propagation paths resulting in multiple replicas of the transmitted signal at the receiver. The receiver will exploit the multiple received replicas of the signal to improve the performance of the detection process. As an alternative, each antenna can be used to transmit independent data over the propagation paths to increase the communication data rate over a channel. In other words, the multiple antenna system can be viewed as having multiple spatial channels over which independent information is propagating to increase the data rate [7]. Fortunately, these performance and rate gains are not mutually exclusive, but there is a fundamental trade-off between them, known as the diversity-multiplexing trade-off [7, 8].

The maximum achievable data rate at which a system can reliably communicate with an arbitrarily small probability of error is known as the channel capacity [9]. In order to describe the channel capacity of a MIMO system, the channel model will be first defined. The considered model employs M transmit and N receive antennas. The complex attenuations of the channels between each pair of transmit and receive antennas can be represented by an $M \times N$ channel realization matrix H. If the channel is constant for blocks of T channel uses, and if the transmitted signal is represented by a $T \times M$ matrix S, the received signal is a $T \times N$ matrix Y given by,

$$Y = SH + V, \tag{1.1}$$

where V is a $T \times N$ matrix that represents the additive noise. This channel model will be described in more detail in Section 2.2. The channel capacity is equal to the maximum mutual information $I(S;Y)$ between the input S and the output Y of the channel over the distribution of the input signals [9]. In this thesis, we consider block-fading channels in which the channel realization H remains constant for a block of channel uses then assumes an independent realization for the next block. Furthermore, we assume that the latency constraints are loose so that the information can be coded across many blocks. For such a scenario, the maximum data rate that can be reliably communicated is bounded by the ergodic capacity of the channel [10]. The value of the ergodic capacity is dependent on the amount of information about the state of the channel that the transmitter and receiver possess. In this thesis, two models that have received considerable attention are described; the coherent and non-coherent models [10, 11]. In both models, the transmitter has no knowledge of the channel.

The coherent communication model assumes that the receiver has perfect knowledge of the channel state information (CSI) [10]. In practice, the CSI can be acquired by transmitting pilot symbols known to the receiver, from which the receiver can

estimate the channel. This strategy is based on the assumption that the length of the block over which the channel is assumed to be constant is long enough for the time and power used to acquire this CSI to be deemed negligible. Since the receiver is assumed to have obtained perfect knowledge of the channel realization H, the coherent system model can be viewed as a channel with two outputs the received signal, Y, and the channel realization, H. Therefore, the ergodic capacity of the coherent model is given by,

$$C_{\text{coherent}} = \max_{p(S)} E_H \{I(S;Y,H)\} \tag{1.2}$$

$$= \max_{p(S)} E_H \{I(S;Y|H)\}, \tag{1.3}$$

where $p(S)$ is the distribution of the input signals, $I(S;Y,H)$ is the mutual information between the channel input S and the channel output pair Y and H, $I(S;Y|H)$ is the mutual information for a given channel realization H and $E_H\{\cdot\}$ denotes the expectation over H. The derivation of (1.3) exploits the fact that $I(S;H) = 0$; i.e., the channel input S and the channel realization H are independent, since H is not known at the transmitter.

An alternative communication model that appropriately accounts for the resources that have to be expended to acquire the CSI is the non-coherent model. In this model, the receiver does not have any CSI. The ergodic capacity of this model can be written as,

$$C_{\text{non-coherent}} = \max_{p(S)} E_H \{I(S;Y)\}. \tag{1.4}$$

Explicitly computing the capacity of a non-coherent system remains an open problem [11]. However, in [12], the non-coherent ergodic capacity of a MIMO system at high signal to noise ratios (SNRs) was computed. While this will be described in more detail in Section 2.2, it was shown in [12] that channel symbols in the form of isotropically distributed unitary matrices achieve the non-coherent ergodic capacity

at high SNRs. It was also shown in [12] that the transmitted information is conveyed in the subspaces spanned by these unitary matrices rather than the matrices themselves, and that each subspace can be represented as constellation point on a Grassmann manifold. This communication scheme will be the focus of Chapter 2.

1.3 Literature Review

The goal of this thesis is to develop a computationally efficient scheme that is capable of communicating reliably at data rates close to the ergodic capacity of a non-coherent MIMO system at high SNRs. A common approach to achieving reliable communication at high rates is by combining a highly structured "outer" code with an "inner" signalling scheme that mimics the capacity achieving distribution. Two well-known generic forms of that approach are bit-interleaved coded modulation (BICM) [13, 14] with iterative demapping and decoding (IDD) [15, 16], and multilevel coded (MLC) modulation [17, 18]. In the BICM-IDD scheme, at the transmitter, the input data are encoded bits using an "outer" code. The encoded bits are scrambled using a pseudo-random interleaver and then consecutively mapped onto a constellation. At the receiver, the demapping of encoded bits from the received signal and the decoding of these bits are performed separately while iteratively exchanging soft information to approximate the optimum joint demapper/decoder [15, 16]. Although BICM-IDD is not capacity achieving, it simplifies the implementation of the both the transmitter and the receiver, while still offering the potential for reliable communication close to the capacity limit [16]. In MLC modulation, the input data is demultiplexed into multiple data streams using a serial-to-parallel converter. The data streams are then independently encoded at different data rates. The rate of each of these data streams is determined via a chain rule decomposition of the mutual information between the input data bits and the received signal [18, 19]. The encoded bits of the parallel data

streams are consecutively mapped to a signalling constellation of a suitable size and then transmitted over the channel. Using the insight generated by the chain rule decomposition of the mutual information, the receiver uses a multistage decoder that detects the data streams in a successive fashion. In particular, the receiver begins by detecting one stream. Assuming that this stream is correctly detected, the receiver uses the recovered data to subsequently detect a second stream. The receiver continues until all data streams have been detected. Although MLC is capacity achieving, implementing MLC schemes at high data rates can be computationally expensive due to the associated complexity of the multistage decoding technique [19].

The focus of this thesis is on applying the principles of BICM-IDD to the non-coherent MIMO channel. Several systems have been developed in [20–23] in which BICM-IDD was applied to the non-coherent channel using various unitary signalling schemes. In order to describe the systems developed in [20–23], different unitary constellation design techniques will first be addressed.

A key aspect in designing a signalling constellation is to determine an appropriate metric that quantifies the distance between the constellation points. Several constellation design techniques using different distance metrics have been proposed for the non-coherent MIMO channel. For example, in [24] the asymptotic union bound on the pairwise error probability was proposed as a distance metric. However, this bound can be loose and this can be reflected in the constellation; i.e., constellations that are deemed "good" from this bound's perspective might not perform well in practice. Another constellation design technique was proposed in [25] that uses an unstructured surrogate-based optimization technique to generate the constellation. This approach adopted a mathematically convenient metric called the chordal Frobenius distance [26]. Although this distance metric is mathematically convenient, it does not necessary guarantee good performance [27]. Another approach was proposed in [27] in which a metric called the chordal Frobenius norm was used. This norm was shown to

quantify the perturbation in the subspace spanned by the transmitted signal induced by the noise at the receiver [27]. Using this metric, a practical greedy algorithm to design the unitary matrices constellation was also proposed in [27].

The BICM-IDD scheme using unitary signalling developed in [20, 22, 23] uses the distance metric and constellation design developed in [25, 28] and uses full constellation demapping which can be computationally expensive especially at high data rates. In [21], an EXIT chart analysis was performed on the BICM-IDD scheme that uses the unitary constellations designed in [28] and full constellation demapping. In BICM-IDD schemes, the encoded bits are mapped onto a constellation of unitary matrices. In the BICM-IDD schemes developed in [20–22], a pseudo-random mapping technique was used to label the constellations of unitary matrices. However, in [23] a mapping technique was proposed in which optimization techniques were used to minimize a cost function based on the pairwise error probability. The developed mapping performs well in the case where the channel variations between two consecutive blocks can be considered negligible. However, in the considered block-fading model this optimized mapping does not show any improvement in performance over pseudo-random mapping [23]. In addition to employing a simple mapping technique that exploits the structure of the unitary constellation, the proposed scheme also differs in the choice of the distance metric and the constellation design technique [27]. In fact, it has already been shown in [27] that for an uncoded scheme the chosen unitary constellation performs significantly better than the constellations developed in [25, 28]. In addition to the differences in the signalling scheme, the proposed receiver is unique at its use of an efficient list-based demapping technique that significantly reduces the complexity of the detection process.

In addition to unitary signalling [12, 29], there are other candidate approaches to non-coherent communication, such as differential schemes [29–31] and training techniques [3, 32, 33]. Differential schemes are based on the assumption that the channel

variations between two consecutive blocks is negligible. However, this assumption is not justifiable under the block-fading channel model used in this thesis, because the channel takes on independent values for each block. In contrast, training-based schemes are appropriate for the block-fading channel model. In the simplest form of such schemes, the channel is first estimated using the training sequence and then the system communicates coherently using the estimated channel as if it was the true channel. More sophisticated schemes involve joint or iterative estimation of the channel and detection of the data [34, 35], but such schemes have a substantially larger computational cost.

In this thesis, the performance of the proposed non-coherent communication scheme will be compared to a corresponding training-based scheme, and it will be shown that the proposed scheme provides better performance than the training based scheme at high data rates.

1.4 Contributions

The focus of this thesis is to construct a scheme that can communicate reliably at data rates close to the ergodic capacity of a non-coherent MIMO system at high SNRs. The proposed scheme applies the principles of BICM-IDD [15,16] to the non-coherent capacity achieving signalling scheme, which is in the form of isotropically distributed unitary matrices [12]. While others have developed BICM-IDD schemes for the non-coherent MIMO channel using various unitary signalling schemes (e.g., [20–23]), the proposed scheme uses a different unitary signalling scheme that more closely mimics the distribution that achieves the capacity at high SNRs. In addition to employing a simple mapping technique that exploits the structure of the unitary constellation, the proposed scheme also employs an efficient list demapping scheme that significantly reduces the computational load at the receiver.

Before applying the principles of BICM-IDD to a high rate communication scheme, one needs to design a constellation of unitary matrices that mimics the capacity achieving distribution. The complexity of directly generating such constellation increases exponentially with the data rate. An alternative approach is through applying a greedy algorithm that generates a well-spaced constellation of unitary matrices by recursively adding one constellation matrix at a time [27]. After designing such constellation, the transmitter of the BICM-IDD scheme maps consecutive blocks of interleaved coded bits onto the designed constellation.

Since efficient deterministic methods for mapping the bits onto such constellation are not yet known, we will use the structure of the Grassmannian constellation to guide the construction of a constellation mapper that maps consecutive blocks of bits onto the constellation points.

The receiver of the BICM-IDD scheme iteratively exchanges soft information between the demapper and decoder to approximate the prohibitively complex jointly optimal receiver [14–16]. However, examining all the constellation points in the demapper of the BICM-IDD scheme is still computationally expensive. To alleviate such a computational burden, one might choose to examine only a subset of the constellation points. In order for the demapper to perform well, the candidate list that represents the subset of the constellation points has to be carefully chosen by exploiting the structure of the Grassmannian constellation and the received signal. The proposed list demapper is based on the recently developed reduced-search detector for uncoded Grassmannian constellations [36,37]. A weakness in the initial list-based demapper is that membership of the candidate list depends only on the channel output without any input from the decoder of the BICM-IDD scheme. In other words, a constellation point that is deemed by the decoder to have a large likelihood might not be a member of the demapper's candidate list. Thus, we propose to incorporate the decoder's information in the demapper by allowing the decoder to augment the

demapper's candidate list. This augmentation results in a significant improvement in performance.

In summary, the proposed scheme applies BICM-IDD to the distance metric and the Grassmannian constellations developed in [27] to generate a coded system that communicate reliably at data rates close to the non-coherent ergodic capacity. A mapping technique that exploits the Grassmannian geometry is proposed. In addition, a computationally-efficient list demapping technique that incorporate the decoder's information in the candidate list is developed. Finally, the performance of the proposed non-coherent BICM-IDD scheme using unitary signalling is compared to a corresponding BICM-IDD training-based scheme [3, 12, 32, 33]. It will be shown, via simulation, that the proposed scheme can provide better performance at high data rates.

1.5 Thesis Outline

The core contribution of this thesis is the development of a coded scheme that operates reliably at data rates close to the ergodic capacity of the channel, using the principles of BICM-IDD [14–16]. In the development of this scheme, we encountered several challenges, most of which were consequences of the non-Euclidean geometry of the capacity achieving signalling scheme. The background to these challenges is described in Chapters 2–4 and in Chapter 5 this scheme is constructed.

Chapter 2 will begin by describing the principles of non-coherent MIMO communication and the corresponding channel model. Since the capacity-achieving signals at high SNRs can be represented as points on a Grassmann manifold [12], we will describe several constellation design techniques that mimic this distribution. Since the demapping process is usually the computational bottle neck of high-rate MIMO communication systems, we will also describe an efficient reduced-search list-based

detector for uncoded signalling over a non-coherent channel that exploits the non-Euclidean structure of the Grassmannian constellations [36, 37]. Finally, we will discuss alternative techniques for non-coherent communication.

Chapter 3 will focus on the principles of the bit-interleaved coded modulation (BICM) scheme described in [14]. Applying the BICM scheme to a MIMO communication system results in a system that consists of both a space-time inner constellation and an outer code that introduces a structure between the transmitted blocks (e.g., a convolutional or turbo code). Since the optimal joint detection and decoding of a BICM scheme is computationally infeasible for large data blocks, we will describe the principles of a sub-optimal technique described in [15, 16] that iteratively exchanges soft information between the detector and the decoder. Finally, the functionality of the BICM scheme with iterative detection and decoding (IDD) will be illustrated for a coherent MIMO system.

Chapter 4 will provide an overview of turbo codes, which will be used as the outer code for the proposed non-coherent BICM-IDD scheme. We will focus on the commonly used class of turbo codes that is constructed by the parallel interleaved concatenation of recursive systematic convolutional codes. Such codes can be simply implemented to allow the BICM-IDD scheme to operate at data rates close to the capacity limit [38]. We will discuss the functionality of the building blocks of both the turbo encoder and its corresponding decoder. Turbo decoding is performed through the iterative exchange of soft information between the decoding blocks (using, for example, the BCJR algorithm [39]). This decoding technique will also be described in detail.

Chapter 5 presents the main contributions of the thesis. The core contribution is to apply the principles of BICM-IDD [14–16] to the non-coherent MIMO channel. Constructing such scheme can be challenging due to the non-Euclidean geometry of the optimal non-coherent space-time signals. In this chapter, we will exploit

the Grassmannian structure to develop a constellation mapper for the transmitter of the BICM scheme that maps consecutive blocks of bits onto the Grassmannian constellation. The constellation points are then transmitted through the described block-fading channel. In constructing the receiver it is observed that examining all the unitary matrices at the receiver is computationally expensive, and hence insight from the geometry of the signalling scheme is used to develop an efficient list-based demapping algorithm that substantially reduces the computational complexity of the receiver while incurring some degradation in performance (around 1.75 dB in SNR at a bit error rate (BER) of 10^{-4}). This demapper is based on a recently developed reduced-search detector for uncoded Grassmannian constellations [27, 37]. Furthermore, we propose a method by which the decoder can augment the list used by the demapper and we demonstrate that this feature renders the performance degradation of the efficient demapper negligible (around 0.2 dB in SNR). Finally, the performance of the proposed scheme will be compared to that of a corresponding training-based BICM-IDD scheme.

Chapter 6 concludes the thesis and provides few directions that can be pursued for future research, with the goal of making further improvements to the performance of the non-coherent MIMO BICM-IDD system.

Chapter 2

Non-Coherent MIMO Communication

2.1 Introduction

As discussed in Chapter 1, the use of multiple antennas at the transmitter and receiver of a wireless communication system offers the potential for reliable communication at high data rates. In a standard model for such a multiple-input-multiple-output (MIMO) communication system, the receiver is assumed to have complete Channel State Information (CSI) *a priori* and is able to use this information to detect the transmitted signals. This model is based on the assumption that the coherence time is long enough for the time used to acquire this CSI to be deemed negligible. Such model is referred to as being coherent [10]. However, this model does not consider the communication resources that would have to be expended in order for the receiver to acquire the CSI. An arguably more realistic framework that allows for these resources to be appropriately accounted for is one in which neither the transmitter nor the receiver has any *a priori* CSI. Such a communication model is often referred to as being non-coherent. Despite the absence of the *a priori* CSI, MIMO non-coherent

communication systems can offer reliable data transmission at high rates [12]. In fact, it was shown in [12] that at high SNRs the non-coherent ergodic channel capacity approaches the capacity of a corresponding coherent model as the coherence time increases.

Explicitly computing the capacity of a non-coherent system remains an open problem [11]. However, in [12], the high SNR non-coherent ergodic capacity of a frequency-flat richly-scattered block-fading MIMO channel was computed. This computation revealed several fundamental relationships between the variables of the non-coherent MIMO system. For example, it was shown that increasing the number of transmit antennas beyond the number of receive antennas will not result in any gain in the achievable rate [12]. In addition, it was shown that the high-SNR capacity achieving input signals are in the form of isotropically distributed unitary matrices [12, 29, 40]. In contrast, in the coherent case a vector version of the conventional Gaussian signalling strategy suffices to achieve the ergodic capacity [10]. This difference reduces the extent to which insight can be transferred from the coherent case to the non-coherent case. However, some insight into the non-coherent case can be obtained by observing that when we communicate non-coherently using unitary matrices, the information is conveyed by the subspace spanned by the transmitted unitary matrices rather than the matrices themselves, and these subspaces can be represented by points on a compact Grassmann manifold [12]. It was shown in [12,25] that the design of the isotropically distributed unitary matrices is identified with constellation points that are well-spaced on a Grassmann manifold. The concepts of this non-coherent MIMO scheme will be described in more detail in the next section.

Despite the insight that is gained from the Grassmannian geometry, the implementation of the optimal non-coherent signalling scheme remains challenging. A major challenge lies in the design of a constellation that mimics the capacity achieving signal distribution (which is in the form of isotropically distributed unitary matrices).

The complexity associated with directly generating such constellation grows rapidly with the constellation size. Thus, an alternative constellation design technique that is based on a greedy algorithm was proposed in [27]. Another challenge lies in the detection process where the optimal maximum likelihood (ML) detection can be computationally infeasible for large constellations at high data rates. Thus, in order to reduce the complexity of the detection, a reduced search non-coherent detector that exploits the structure of the Grassmannian manifold and the received signal was introduced in [27,37].

This chapter will present an overview of MIMO non-coherent communication. Section 2.2 will introduce the channel model, and then Section 2.3 will address a practical constellation design technique. In Section 2.4, the ML detection procedure will be discussed and a computationally efficient non-coherent reduced search detector [27] will be described. Finally, Section 2.5 will discuss alternative techniques that could be used to communicate over a non-coherent channel.

2.2 System Model

We consider a system in which information is communicated from M transmit antennas to N receive antennas over a frequency-flat richly-scattered block-fading channel of coherence time T_c channel uses. This system model is appropriate for time-division multiple access (TDMA) or frequency hopping systems [3], as described in Section 1.1. We will denote the signalling interval (in channel uses) by T, and we will choose $T \leq T_c$ so that we can model the channel as being constant over the signalling interval. The vector of signals transmitted at each channel use will be denoted by the rows of a $T \times M$ matrix Q_X. The $T \times N$ received signal matrix Y can then be written as

$$Y = Q_X H + \sqrt{\frac{M}{\rho T}} V, \tag{2.1}$$

where H is an $M \times N$ channel matrix whose entries are drawn independently from the standard complex Gaussian distribution $\mathcal{CN}(0,1)$, V is the $T \times N$ matrix representing the additive noise whose entries are also drawn independently from $\mathcal{CN}(0,1)$, and ρ is the signal to noise ratio. (The scaling in (2.1) is performed so that ρ is independent of both M and T.)

For the non-coherent MIMO system, the capacity achieving input signals for high-SNR operation are isotropically distributed $T \times M$ (tall) unitary matrices Q_X. These tall matrices Q_X are unitary in the sense that $Q_X^\dagger Q_X = I_M$, where $(\cdot)^\dagger$ represents the Hermitian transpose of a matrix, and they are isotropically distributed in the sense that $P(Q_X) = P(UQ_X)$ for any $T \times T$ unitary matrix U, [12,40]. When the transmitted $T \times M$ matrix Q_X is right multiplied by the $M \times N$ channel matrix H as shown in (2.1), the basis vectors of the M-dimensional subspace are rotated and scaled. However, since the receiver has no channel information, this rotation and scaling of the basis vectors cannot be detected. However, the subspace spanned by Q_X remains unchanged [12]. Therefore, information is conveyed in the M-dimensional subspace spanned by Q_X [12]. This comes in contrast with the coherent scenario in which the receiver has perfect knowledge of the channel H, and therefore the information-carrying object is the transmitted matrix Q_X itself. Since in the non-coherent scheme only the subspace spanned by Q_X is detectable, all unitary matrices that span the same M-dimensional subspace are equivalent from the perspective of the high-SNR non-coherent MIMO communication system. That is, if a $T \times M$ unitary matrix Q_X spans a certain subspace Φ, then all matrices $Q_X P$ for any $M \times M$ unitary matrix P will span the same subspace Φ and are considered equivalent [12]. The subspace spanned by each matrix Q_X can be represented by a single "constellation" point on a compact Grassmann manifold $\mathbb{G}_M(\mathbb{C}^T)$ [12], where $\mathbb{G}_M(\mathbb{C}^T)$ is the set of all M-dimensional subspaces that span a T-dimensional Euclidean space.

Explicitly computing the mutual information between the input Q_X and output Y

of a non-coherent system remains an open problem [11]. However, the non-coherent ergodic capacity at high SNRs was computed in [12]. This computation used the fact that the mutual information $I(Q_X;Y) = h(Y) - h(Y|Q_X)$, where $h(\cdot)$ is the entropy and it was based on the assumption that the SNR is high so that the entropy of the channel output $h(Y) \approx h(Q_X H)$; meaning that the entropy of the noise is negligible when compared to the entropy of $Q_X H$; i.e., $h(V) \ll h(Q_X H)$. It was shown in [12] that for a non-coherent MIMO system with $N \geq M$ and $T \geq M + N$, the ergodic capacity can be written as

$$C(\rho) = M(1 - \frac{M}{T}) \log_2(\rho) + c_{M,N} + o(1), \tag{2.2}$$

where $o(1)$ is the Landau symbol which becomes insignificant as the SNR goes to infinity [12]. The constant $c_{M,N}$ is given by

$$c_{M,N} = \frac{1}{T} \log_2(|\mathbb{G}_M(\mathbb{C}^T)|) + M(1 - \frac{M}{T}) \log_2(\frac{T}{\pi e M}) + (1 - \frac{M}{T}) \sum_{i=N-M+1}^{N} E\{\log_2(\chi^2_{2i})\}, \tag{2.3}$$

where χ^2_{2i} is a Chi-square random variable with $2i$ degrees of freedom, and $|\mathbb{G}_M(\mathbb{C}^T)|$ is the volume of the Grassmann manifold $\mathbb{G}_M(\mathbb{C}^T)$, which is given by

$$|\mathbb{G}_M(\mathbb{C}^T)| = \frac{\prod_{i=T-M+1}^{T} \frac{2\pi^i}{(i-1)!}}{\prod_{i=1}^{M} \frac{2\pi^i}{(i-1)!}}. \tag{2.4}$$

It was shown in [12], that increasing the number of transmit antennas M beyond the number of receive antennas N will not result in any increase in capacity. Furthermore, in order to maximize the SNR-dependent term in the non-coherent capacity in (2.2), the appropriate number of transmit antennas is

$$M = \min\{\lfloor T/2 \rfloor, N\}, \tag{2.5}$$

and we will assume that (2.5) is satisfied.

Since the information in a non-coherent MIMO system that operates at high-SNR is conveyed in the subspaces spanned by the unitary matrices Q_X, the task of

the detector is to identify the subspace that was transmitted rather than the actual matrix that generated the subspace. The received $T \times N$ matrix Y, which spans an N-dimensional space, can be decomposed using QR decomposition. That is,

$$Y = Q_Y R_Y,$$

where the columns of Q_Y constitute a basis of the N-dimensional subspace in which the received signal lies, and R_Y consists of the scaling and rotating factors within the subspace. It was shown in [27], that for isotropically distributed unitary matrices Q_X, the unitary component of the received signal Q_Y is independent from the channel H, while R_Y is independent from the transmitted signal Q_X. That is,

$$I(Q_Y; H) = 0, \tag{2.6}$$

$$I(R_Y; Q_X) = 0. \tag{2.7}$$

The proof of this result is based on the fact that since the noise and the transmitted matrices are isotropically distributed, the noise does not couple the information about the channel and the transmitted matrices [27]. Based on (2.6) and (2.7), all the available information about the subspace of Q_X is contained in Q_Y, while R_Y includes all the available information about the channel H. Therefore, the role of the detector is to exploit the subspace spanned by Q_Y and the implicit channel information in R_Y to detect the subspace of the transmitted unitary matrix Q_X (see Section 2.4).

2.3 Constellation Design

As described in the previous section, an input distribution that is isotropically distributed on the space of $T \times M$ unitary matrices achieves the ergodic capacity of the non-coherent MIMO system at high SNRs [12, 29, 40]. The goal of the constellation design is to synthesize a constellation of unitary matrices that mimics this

distribution. It was shown in [12, 25] that the design of the isotropically distributed unitary matrices is identified with constellation points that are uniformly distributed on a Grassmann manifold. Therefore, in order to design such constellation, one could attempt to design a constellation of this form by directly maximizing the mutual distances between the subspaces that the matrices Q_X span; that is,

$$\begin{aligned} &\arg \max_{\{Q_{X_k}\}_{k=1}^{|\mathcal{C}|}} \min_{\substack{1\leq i,j\leq|\mathcal{C}| \\ i\neq j}} d(Q_{X_i}, Q_{X_j}), \\ \text{subject to} \quad &Q_{X_i} \in \mathbb{G}_M(\mathbb{C}^T), \quad \forall i \in \{1, 2, 3, ..., |\mathcal{C}|\}, \end{aligned}$$

where $d(\cdot, \cdot)$ is an appropriate distance metric for a Grassmannian constellation, and $|\mathcal{C}|$ is the number of constellation matrices. The computational cost of this approach increases rapidly with the size of the constellation. Furthermore, there is some debate as how one ought to measure the distances between the subspaces [25, 27, 41]. It was shown in [27] that using the chordal Frobenius norm $d_{CF}(\cdot, \cdot)$ as the distance metric accurately accounts for the subspace perturbation due to the noise at the receiver. The chordal Frobenius norm is defined as the square root of

$$d_{CF}^2(Q_{X_1}, Q_{X_2}) = 2M - 2\text{Tr}(\Sigma_{Q_{X_1}^\dagger Q_{X_2}}), \tag{2.8}$$

where the SVD of $Q_{X_1}^\dagger Q_{X_2} = U_{Q_{X_1}^\dagger Q_{X_2}} \Sigma_{Q_{X_1}^\dagger Q_{X_2}} V_{Q_{X_1}^\dagger Q_{X_2}}^\dagger$. In addition, it was shown that constellations designed using this distance metric provide significantly better performance than the constellations designed using the chordal Frobenius distance in [25,28]. The chordal Frobenius distance is the square root of $M - \text{Tr}(\Sigma_{Q_{X_1}^\dagger Q_{X_2}}^2)$, [26]. Therefore, in this thesis we will choose the constellation designed using the greedy algorithm proposed in [27] in which the chordal Frobenius norm was chosen as the distance metric.

Starting from an arbitrary unitary matrix Q_X the greedy design technique in [27] recursively adds one constellation matrix at a time. Each additional matrix spans a subspace that maximizes the minimum of the distances to all the subspaces that

are spanned by the existing matrices in the constellation; i.e., the approach in [27] attempts to iteratively solve

$$\begin{aligned} Q_{X_i} &= \arg \max_{\{Q|Q^\dagger Q=I\}} \min_{1\le j\le i-1} d_{CF}(Q, Q_{X_j}), \\ &= \arg \min_{\{Q|Q^\dagger Q=I\}} \max_{1\le j\le i-1} \mathrm{Tr}(\Sigma_{Q^\dagger Q_{X_j}}), \quad i = 2,3,...,|\mathcal{C}|. \end{aligned} \tag{2.9}$$

As it stands, (2.9) is still an awkward problem to solve, because the inner maximization problem renders the objective of the minimization problem non-differentiable. In order to use efficient techniques for smooth optimization on the Grassmann manifold, the technique in [27] uses the sum of negative exponentials of the distances to approximate the inner maximization in (2.9). This results in the following smooth optimization problem,

$$Q_{X_i} = \arg \min_{\{Q|Q^\dagger Q=I\}} \sum_{j=1}^{i-1} \exp(\mathrm{Tr}(\Sigma_{Q^\dagger Q_{X_j}}))^2. \tag{2.10}$$

In the case where the signalling interval is twice the number of transmitters, $T = 2M$, the symmetry properties of the underlying Grassmann manifold mean that the constellation can be increased by an orthogonal pair of unitary matrices at each iteration [27]. This property will be exploited in the BICM-IDD non-coherent system in Section 5.2 where binary bits will be mapped to points on the compact Grassmann manifold. More detailed description of the smoothed greedy constellation design technique and a detailed derivation of the choice of distance metric can be found in [27, 36].

2.4 Non-Coherent Detection

In any uncoded non-coherent communication system with unitary signalling, the receiver's role is to detect which one of the "constellation" of subspaces was transmitted.

Maximum likelihood (ML) detection of the subspace involves computing the likelihood of the received signal given the transmitted unitary matrix for all constellation points. In fact, under the standard assumptions that the entries of the channel and noise matrices are drawn independently from the standard complex Gaussian distribution $\mathcal{CN}(0,1)$ and the channel model in (2.1), the likelihood of the received signal given the transmitted unitary matrix is [12,40]

$$p(Y|Q_X) = \frac{\exp\left(-\frac{\rho T}{M}\mathrm{Tr}(Y^\dagger(I_T - \frac{1}{1+M/\rho T}Q_X Q_X^\dagger)Y)\right)}{(\pi M/\rho T)^{TN}(1+\rho T/M)^{MN}}, \tag{2.11}$$

where I_T is the $T \times T$ identity matrix. The ML detector examines all the constellation points in the constellation set $\mathcal{C}$ searching for the constellation point $\hat{Q}_X$ that maximizes $p(Y|Q_X)$ in (2.11). This is equivalent to

$$\hat{Q}_X = \arg\max_{Q_X \in \mathcal{C}} \mathrm{Tr}(Y^\dagger Q_X Q_X^\dagger Y). \tag{2.12}$$

The drawback of ML detection lies in the computational cost of examining all the constellation points in the constellation set. The computational complexity of the ML detector for each received matrix Y is equal to $(TMN + NM)|\mathcal{C}|$ complex multiplications and $[(T-1)MN + (NM-1)]|\mathcal{C}|$ complex additions. This number of computations is based on only computing the diagonal entries of $Y^\dagger Q_X Q_X^\dagger Y$ that is needed to compute the trace. In order to alleviate the computational cost of the detection process, a reduced-search non-coherent list detector was developed in [27,37]. This detector selects a particular subset of candidate constellation points to be examined against the ML metric (2.12).

2.4.1 Reduced-Search Non-Coherent Detector

The reduced search detector for uncoded Grassmannian constellations developed in [27, 37] uses the structure of the Grassmannian constellation and the nature of

the received signal Y to determine a list of candidate constellation points. A distinguishing feature of the list is that its length is implicitly adapted to the channel realization, rather than being fixed *a priori*. This enables the receiver to allocate its computational resources to channel realizations for which the detection problem is "hard". The generation of the list of candidate constellation points is based on the QR decomposition, $Y = Q_Y R_Y$, and the observation [36,37] that all the information available about the transmitted signal, Q_X, is contained in Q_Y, while R_Y contains the available information about the channel, H; c.f., (2.6) and (2.7). To simplify the description of the list generation scheme, in this thesis we will focus on systems in which the number of transmit antennas is the same as the number of receive antennas; i.e., $M = N$. Extensions to the case in which $N > M$ appear in [36,37].

A list of candidate constellation points can be generated as follows: Prior to operation, the detector picks a reference constellation point $Q_{\text{ref},1}$ and builds a look-up table in which the remaining constellation points are sorted according to their distance from the reference point. When a signal matrix Y is received , its QR decomposition is computed and the distance $d(Q_Y, Q_{\text{ref},1})$ is measured. All constellation points which are at "about the same distance" from the reference point as Q_Y are included in the candidate list. More specifically, the channel information implicit in R_Y is used to generate two values, A_Y and B_Y, which are used to define the width of a "strap" on the Grassmannian manifold that contains Q_Y; see Figure 2.1 for an illustration. The detector's list is defined to be all those constellation points that lie within the strap; i.e.,

$$\mathcal{L}(Y, Q_{\text{ref},1}) = \{Q_X | A_Y \leq d(Q_X, Q_{\text{ref},1}) - d(Q_Y, Q_{\text{ref},1}) < B_Y\}. \tag{2.13}$$

In order to operate with shorter lists, the look-up table in the detector can be augmented to include distances from other reference points $Q_{\text{ref},j}$. The detector will augment the look-up table with the distances between all the constellation points and each additional reference point $Q_{\text{ref},j}$. The look-up table is computed only once and

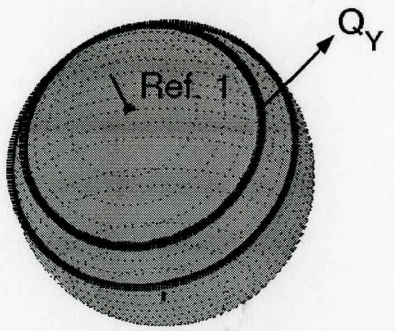


Figure 2.1: A "strap" on the Grassmann manifold $\mathbb{G}_1(\mathbb{R}^3)$ that contains Q_Y. (This figure originally appeared in [36] and is used with permission.)

stored at the receiver. In the case of multiple reference points, only those constellation points that lie in the intersection of the "straps" of each reference point are assigned to the list; see, e.g., Figure 2.2. That is, the list can be described as

$$\mathcal{L}' = \{Q_X | Q_X \in \bigcap_j \mathcal{L}(Y, Q_{\text{ref},j})\}. \tag{2.14}$$

Having established the list, the detector then selects the constellation point in the candidate list that maximizes the maximum likelihood metric (2.11). That is, the detector chooses

$$\hat{Q}_X = \arg \max_{Q_X \in \mathcal{L}'} \text{Tr}(Y^\dagger Q_X Q_X^\dagger Y). \tag{2.15}$$

The choice of A_Y and B_Y in (2.13) involves a trade-off between the length of the list and the probability that the transmitted constellation point is not in the list. A good choice for A_Y and B_Y would be the one that minimizes the width of the "strap" $|A_Y - B_Y|$ while ensuring that the probability that the correct codeword is not in the detector's list, $\mathcal{L}'$, is less than a small threshold δ [36,37]. This can be written as,

$$(A_Y, B_Y) = \arg \inf_{\{(A,B)|P(Q_X \notin \mathcal{L}'|Q_X,Y)<\delta\}} |A - B|. \tag{2.16}$$

Computing the probability $P(Q_X \notin \mathcal{L}'|Q_X, Y)$ depends on the choice of the reference point and is rather complicated. As an alternative, one can use the Chebychev inequality to find a lower bound on the width of the strap $|A_Y - B_Y|$ that is required for

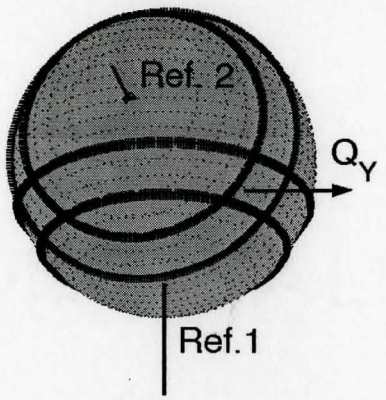


Figure 2.2: The intersection of two "straps" on the Grassmannian manifold $\mathbb{G}_1(\mathbb{R}^3)$. (This figure originally appeared in [36] and is used with permission.)

the probability of "missing" the transmitted constellation point to be below δ [36,37].

A drawback of this detector is the memory used to store all the distances from each reference point $Q_{\text{ref},j}$ in the look-up table. A reasonable solution for such problem is to quantize the distances from the reference points to the constellation points [36,37]. Such approach can result in a valuable reduction in the memory required by the detector.

2.5 Alternative Techniques for Non-Coherent Communication

There are alternative techniques for communicating over a non-coherent MIMO channel other than the capacity achieving Grassmannian signalling technique which has been, so far, the focus of this chapter. In particular, for systems in which one can assume that the channel variations between two consecutive blocks are negligible, differential schemes are strong candidates; e.g., [29–31]. However, this assumption is not justifiable under the block-fading channel model used in this thesis, because the channel takes on independent values for each block.

Another class of non-coherent MIMO communication signalling strategies that can be implemented under the block-fading model are the training-based techniques [3, 12, 32, 33]. Training-based techniques split the signalling interval T into two intervals. The first interval is a training phase, T_τ, in which pilot symbols, that are known to the receiver are transmitted. Using these pilot symbols, the receiver generates an estimate $\hat{H}$ of the channel matrix. In the second interval, T_d, the receiver uses the estimated channel matrix $\hat{H}$ as the true channel to detect the transmitted data coherently [3]. (The training and coherent communication intervals need not be contiguous. Furthermore, joint channel estimation and decoding can be attempted [35].) There is a trade-off between the time used to train the receiver and the time available for data transmission. If the training time, T_τ, is short, the receiver's estimate of the channel $\hat{H}$ can be rather inaccurate and if T_τ is too long, there is little time for data transmission before the channel changes. An appropriate value for the training interval T_τ will be suggested below.

We consider a channel model with M transmit and N receive antennas. The received $N \times T$ matrix Y can be represented in the form of,[1]

$$Y = HS + \sqrt{\frac{M}{\rho}}V, \tag{2.17}$$

where H is the $N \times M$ channel matrix whose entries are drawn independently from $\mathcal{CN}(0,1)$, S is the $M \times T$ space-time codeword that represents the transmitted symbols, V is the $N \times T$ matrix representing the additive noise whose entries are also drawn independently from $\mathcal{CN}(0,1)$, and the signal to noise ratio is given by ρ.

Unlike the unitary signalling technique, in which the constellation is a set of isotropically distributed unitary matrices, the transmitted space-time signal S in the training-based scheme can be constructed from a (widely) linear combination of symbols s from conventional scalar constellations, such as Quadrature amplitude

[1] This form is the transpose of the form in (2.1) with appropriate scaling.

modulation (QAM) or phase shift keying (PSK); e.g., [42]

The training scheme uses pilot symbols known to the receiver to estimate the channel $\hat{H}$. There are different methods to estimate the channel [35]. One simple technique is to transmit the pilot symbols S_τ and then process the received matrix Y_τ to estimate the channel, using either the ML estimate or the linear minimum mean-square error (LMMSE) estimate [3]. These estimates are given by,

$$\hat{H}_{ML} = \sqrt{\frac{M}{\rho}}(S_\tau^\dagger S_\tau)^{-1} S_\tau Y_\tau, \tag{2.18}$$

$$\hat{H}_{LMMSE} = \sqrt{\frac{M}{\rho}}\left(\frac{M}{\rho} I_M + S_\tau^\dagger S_\tau\right)^{-1} S_\tau^\dagger Y_\tau, \tag{2.19}$$

respectively, where I_M is the $M \times M$ identity matrix. In order to acquire a meaningful estimate of the channel H, the number of measurements has to be at least as many as the number of unknowns [3]. That is, the number of elements in the received signal Y_τ is at least as many as the number of unknowns in the channel H; i.e., $NT_\tau \geq NM$. Hence, the number of channel uses T_τ needed to estimate a sufficiently accurate channel matrix is at least as large as the number of transmit antennas M, i.e., $T_\tau \geq M$. In the second phase of the training scheme, the transmitter and the receiver will communicate coherently, with the receiver using the estimated channel, $\hat{H}$, as if it was the true channel to detect the transmitted signal. In the next subsection, we will provide a simple example of a 2×2 communication system in which ML detection of the space-time signal S collapses to scalar detection of the elements s in the space-time matrix S.

2.5.1 Example: 2×2 MIMO Communication System

In this example, we will start by describing a signalling scheme for the coherent phase, then training will be applied for this scheme. For a MIMO system in which the number of transmit and receive antennas $M = N = 2$, a convenient technique to communicate

is the Alamouti scheme [43]. In the Alamouti scheme, the transmitted signal, S, possesses an orthogonal structure. This structure allows complex transmitted symbols to be efficiently decoupled at the receiver using linear preprocessing without incurring any penalty in performance [43].

The Alamouti scheme will be described for a 2×2 system with coherence time $T = 2$. Let $\{s_i\}$ denote the complex data symbols. In this scheme, two symbols are transmitted in each block (e.g., s_0 and s_1) which are drawn from a standard constellation. In the Alamouti scheme, the $M \times T$ space time codeword S for the k^{th} block can be written as,

$$S_k = \begin{bmatrix} s_0 & -s_1^* \\ s_1 & s_0^* \end{bmatrix}, \tag{2.20}$$

where $(\cdot)^*$ denotes the complex conjugate. Each row in the S matrix represents a transmit antenna and each column represents a specific time slot. The received matrix Y can be written as,

$$Y = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} s_0 & -s_1^* \\ s_1 & s_0^* \end{bmatrix} + \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}. \tag{2.21}$$

The received matrix Y in (2.21) can be written in a vectorized form where the columns of the matrices are stacked in a tall column,

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{21} \\ y_{12}^* \\ y_{22}^* \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \\ h_{12}^* & -h_{11}^* \\ h_{22}^* & -h_{21}^* \end{bmatrix} \begin{bmatrix} s_0 \\ s_1 \end{bmatrix} + \begin{bmatrix} v_{11} \\ v_{21} \\ v_{12}^* \\ v_{22}^* \end{bmatrix} = \mathcal{H}\mathbf{s} + \mathbf{v}. \tag{2.22}$$

In a coherent channel model, the receiver has access to the channel matrix H. Since the columns of the matrix $\mathcal{H}$ are orthogonal, thus $\mathcal{H}^\dagger\mathcal{H} = \|H\|^2 I$, where $(\cdot)^\dagger$ is the Hermitian transpose and $\|\cdot\|$ is the Frobenius norm. The preprocessed received signal

can be written as,

$$\begin{bmatrix} r_0 \\ r_1 \end{bmatrix} = \mathcal{H}^{\dagger}\mathbf{y} = \|H\|^2 \begin{bmatrix} s_0 \\ s_1 \end{bmatrix} + \mathcal{H}^{\dagger}\mathbf{v}, \tag{2.23}$$

where r_0 and r_1 are the linearly preprocessed received symbols. Thus ML detection can be simply performed by comparing each received symbol r_0 and r_1 to each symbol in the QAM constellation. The detected complex symbols $\hat{s}_0$ and $\hat{s}_1$ can be represented as,

$$\hat{s}_0 = \arg\min_{s} \left\| \tfrac{1}{\|H\|^2} r_0 - s \right\|, \qquad \hat{s}_1 = \arg\min_{s} \left\| \tfrac{1}{\|H\|^2} r_1 - s \right\|. \tag{2.24}$$

As shown in (2.24), ML detection reduces to scalar detection through linear preprocessing of the received signal. In addition, in the Alamouti scheme two versions of each symbol are transmitted on a different antenna at different time instants, as shown in (2.20). The receiver exploits both replicas of the transmitted signal to obtain the diversity advantage described in Section 1.2. As defined in [7], the diversity gain D is the rate at which the ML error probability P_e decays for large SNRs. That is,

$$\lim_{\mathrm{SNR}\to\infty} \frac{\log P_e(\mathrm{SNR})}{\log(\mathrm{SNR})} = -D. \tag{2.25}$$

It was shown in [42], that the Alamouti scheme achieves the maximum diversity gain for a 2×2 MIMO system; i.e., $D = 4$.

Now, lets consider the training-based technique, in which the channel H is not known *a priori* at the receiver. The training-based technique will use pilot symbols known at the receiver to estimate the channel $\hat{H}$ using (2.18) or (2.19)[2]. The Alamouti scheme can then use the estimated channel as if it is the true channel to communicate coherently over the MIMO channel.

In this thesis, this training-based scheme will be used to communicate over a non-coherent MIMO channel in a bit-interleaved coded modulation with iterative demapping and decoding (BICM-IDD) system. The performance of this system will

[2]In this thesis, we used the identity matrix as the pilot space-time matrix to estimate the channel.

be compared to the proposed BICM-IDD system that uses the capacity achieving unitary signalling scheme for non-coherent MIMO communication. The next chapter will describe the principles of BICM-IDD.

Chapter 3

Bit-Interleaved Coded Modulation

3.1 Introduction

In the previous chapter, we discussed the principles of non-coherent MIMO communication systems. The focus of this thesis is to develop a scheme that communicates reliably at data rates close to the ergodic capacity of the non-coherent MIMO channel. One way is by enveloping the non-coherent MIMO system in a bit-interleaved coded modulation (BICM) scheme. In order to provide a complete exposition, in this chapter we will outline the principles of BICM and the advantages it possesses. We will then illustrate the principles by applying them to a coherent MIMO system.

As discussed in Chapter 1, employing multiple antennas at the transmitter and the receiver enables one to achieve high data rates on a richly-scattered wireless channel. One way to construct a scheme that can communicate reliably at these high data rates is by enveloping the multiple antenna system by an "outer" code that introduces a structure between the transmitted symbols. This structure can be exploited at the receiver to correct errors that may result from noise and fading. In such a scheme, the multiple antenna system transmits matrices which are chosen from a space-time constellation that are interlinked through the embedded structure that is introduced

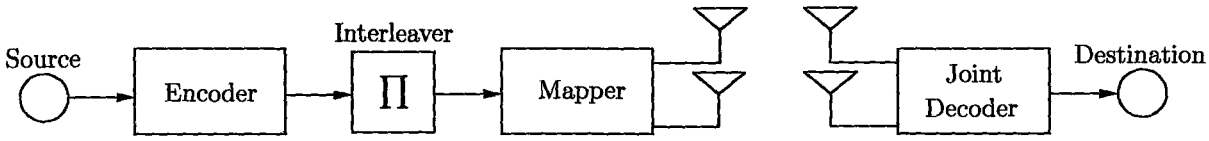


Figure 3.1: Bit-interleaved coded modulation (BICM) scheme with joint decoding.

by the outer encoder. This scheme is referred to as bit-interleaved coded modulation (BICM) [13, 14].

In order to recover the transmitted space-time signals and the information bits in an optimum fashion, one must incorporate both the code structure and the space-time constellation structure into the decoding process [16]. While this joint demapping and decoding scheme shown in Figure 3.1 has the potential of providing attractive performance features, performing maximum likelihood sequence decoding in a BICM scheme involves the examination of all of the codewords, and hence is prohibitively complex [16]. An alternative technique is to perform bit-wise decoding, however this also involves a computational burden that renders it infeasible from a practical point of view. The source of this burden is the need to compute the *a posteriori* probability APP of each bit given all the received data and the outer code constraints. In order to construct a receiver that exploits the structure introduced by the outer code and the space-time constellation, one might choose to separate the demapping and the decoding process [14]; see Figure 3.2. However, coarsely separating the demapping and the decoding can result in a significant degradation in performance [5]. To mitigate this degradation in performance, one might choose to allow the demapper and the decoder to iteratively exchange "soft" information in an attempt to approximate the optimal bit-wise decoder while maintaining reasonable complexity [15, 44]; see Figure 3.3. The demapper and the decoder exchange "soft" information in the form of *a priori* probabilities for several iterations then the final bit values are then generated by performing a hard decision of the "soft" output of the decoder. This

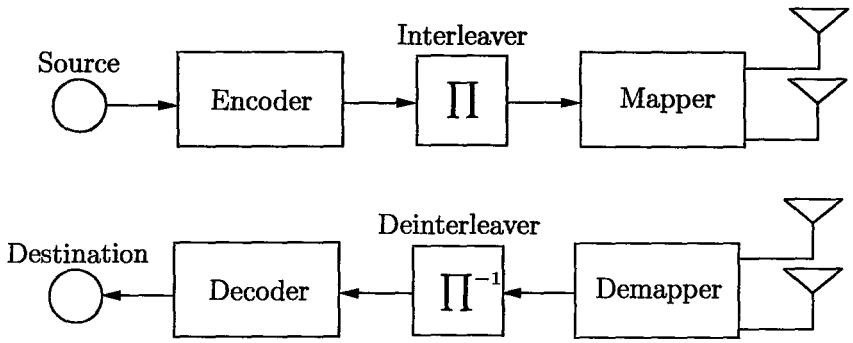


Figure 3.2: Bit-interleaved coded modulation (BICM) scheme.

scheme is referred to as bit-interleaved coded modulation with iterative demapping and decoding (BICM-IDD), and the generic form of a BICM-IDD scheme for a MIMO system is shown in Figure 3.3. In the next section, the principles of the operation of a BICM-IDD scheme for this generic MIMO system will be described. The particular example of a BICM-IDD scheme for a coherent MIMO system with V-BLAST signalling developed in [16] will be considered in Section 3.3.

3.2 Principles of BICM-IDD Scheme

As illustrated in Figure 3.3, the basic framework of a BICM-IDD scheme consists of inner and outer phases. These phases are separated by an interleaver, and are implemented sequentially at the transmitter and iteratively at the receiver [14–16]. The outer phase consists of a binary encoder and its corresponding soft-input soft-output decoder, while the inner phase consists of a constellation mapper at the transmitter that maps blocks of n bits to points on a constellation of size 2^n and a demapper at the receiver that computes the "soft" information of each bit given the received signal [44]. The principles of the transmitter and receiver in the BICM-IDD scheme will be described for a MIMO system in the following sections.

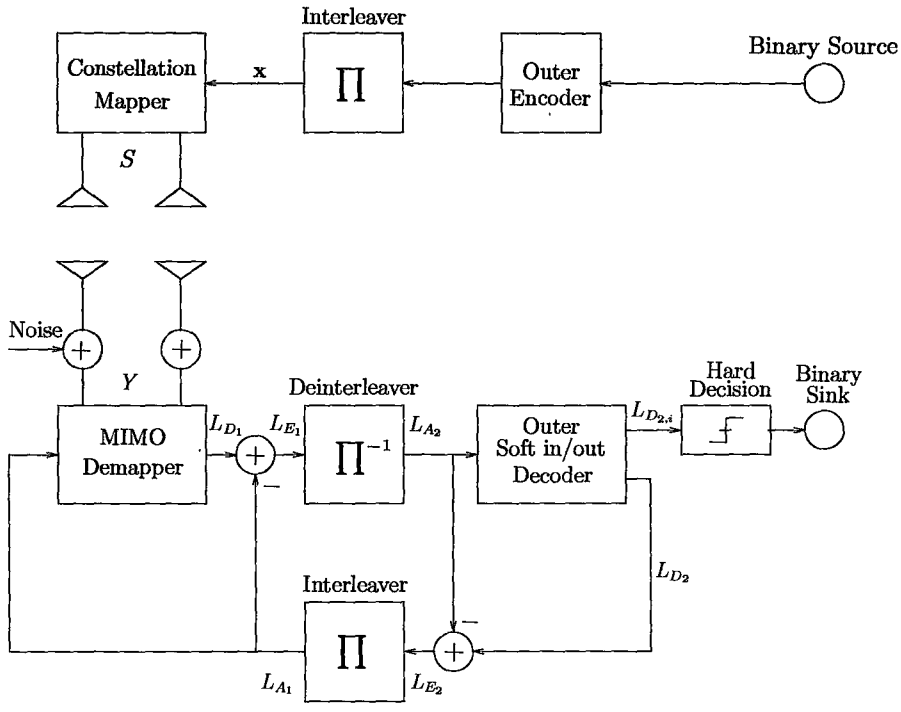


Figure 3.3: A BICM-IDD scheme for a MIMO communication system.

3.2.1 Transmitter

As shown in Figure 3.3, the information bits are passed into an outer encoder (e.g., convolutional or turbo encoder) which will introduce a structure that correlates the coded bits. The coded bits are then scrambled by a pseudo-random interleaver resulting in an interleaved codeword $\mathbf{x}$. The binary elements of $\mathbf{x}$ are then divided into B sequences of n bits, where each sequence $\mathbf{x}^{(b)}$ will be transmitted in one channel use. The interleaver scrambles the bits, so that the n bits in each sequence $\mathbf{x}^{(b)}$ can be considered approximately independent. Each sequence $\mathbf{x}^{(b)}$ will then be mapped onto a space-time matrix symbol $S^{(b)}$ from a constellation set of size 2^n.[1] This matrix can be written as

$$S^{(b)} = \mathcal{M}(\mathbf{x}^{(b)}), \tag{3.1}$$

[1]We will concentrate on systems in which the outer code produces codewords of length Bn, for some integer B.

where $\mathcal{M}(\mathbf{x}^{(b)})$ denotes the mapping of the n-bit $\mathbf{x}^{(b)}$ sequence onto the matrix symbol $S^{(b)}$ from the space-time constellation $\mathcal{C}$.

The space-time symbol $S^{(b)}$ chosen by the block of interleaved coded bits $\mathbf{x}^{(b)}$ are then transmitted through a multiple antenna channel. We consider a multiple antenna system with M transmit antennas and N receive antennas. The received $N \times T$ matrix Y can be written as

$$Y = HS + V, \tag{3.2}$$

where H is the $N \times M$ channel matrix, S is the $M \times T$ transmitted matrix, V is the $N \times T$ additive noise matrix, and T is the signalling block length which is chosen to be no greater than coherence time T_c of the block-fading channel.

3.2.2 Receiver

In order to recover the transmitted signal in an optimum fashion, the receiver would have to jointly compute the likelihood of each bit given all the received blocks $[Y^{(1)}, Y^{(2)}, ..., Y^{(B)}]$ and the outer code constraints [16]. In schemes that operate at rates close to the ergodic capacity of the channel, the blocks of data are usually rather long, rendering optimal joint demapping/decoding computationally infeasible. An alternative sub-optimal approach is through the iterative exchange of soft information between the demapper and the decoder [16].

In the iterative demapping and decoding technique, the demapper and decoder are considered as separate entities, with each using soft information from the other as *a priori* information. The principles of the iterative demapping and decoding algorithm in Figure 3.4 can be described as follows. The demapper observes the received signal Y and the *a priori* soft information L_{A_1} from the decoder. (The soft information will be formally defined below.) The demapper computes the soft information L_{D_1} for each coded bit $x_k^{(b)}$ given the channel observation $Y^{(b)}$, where the k^{th} bit in $\mathbf{x}^{(b)}$

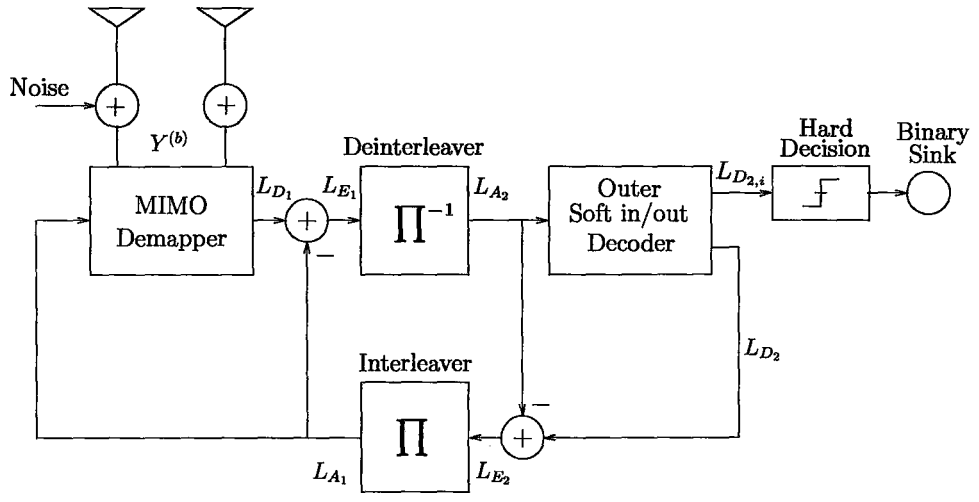


Figure 3.4: The receiver of the BICM-IDD scheme for a MIMO communication system.

is denoted by $x_k^{(b)}$. In order to avoid re-enforcing the existing soft information in the iterative exchange of soft information between the demapper and the decoder, only the extrinsic information is exchanged. The extrinsic information L_{E_1} is obtained by separating the *a priori* soft information from the demapper's output L_{D_1}. The extrinsic information L_{E_1} is deinterleaved to become the *a priori* soft input L_{A_2} to the decoder. The soft-in/soft-out decoder exploits the structure of the code and uses L_{A_2} and all the received blocks $[Y^{(1)}, Y^{(2)}, ..., Y^{(B)}]$ to compute the *a posteriori* probability for each bit which will be used to compute the decoder's soft output L_{D_2}. The soft information L_{D_2} is separated from the *a priori* information L_{A_2} resulting in the extrinsic information L_{E_2}. The extrinsic information L_{E_2} is then interleaved and will act as the *a priori* information L_{A_1} for the demapper; completing an iteration. The demapper and decoder can continue to exchange information in an iterative fashion improving the performance of the system [16].

A convenient representation of the soft information is in the form of log-likelihood

ratios (L-values) [44]. The L-value for a given bit $x_k^{(b)}$ can be written as

$$L_{A_1}(x_k^{(b)}) = \ln \frac{P(x_k^{(b)} = +1)}{P(x_k^{(b)} = -1)}. \tag{3.3}$$

L-values are convenient for iterative decoding since they only require simple subtraction to separate the soft information generated by the demapper or decoder L_D from the *a priori* information L_A resulting in the extrinsic information $L_E = L_D - L_A$, [44].

Maximizing the *a posteriori* probability $P(x_k^{(b)} = +1|Y^{(b)})$ for each bit in the receiver minimizes the chance of having an error in this certain bit. For a received channel observation $Y^{(b)}$, the *a posteriori* L-value for the k^{th} bit in the b^{th} block $x_k^{(b)}$ given the received block $Y^{(b)}$ is

$$L_D(x_k^{(b)}|Y^{(b)}) = \ln \frac{P(x_k^{(b)} = +1|Y^{(b)})}{P(x_k^{(b)} = -1|Y^{(b)})}, \qquad \forall k = 1, 2, ..., n. \tag{3.4}$$

Using Bayes' rule the *a posteriori* L-value in equation (3.4) can be written as [16]

$$L_D(x_k^{(b)}|Y^{(b)}) = \ln \frac{p(Y^{(b)}|x_k = +1)P(x_k = +1)}{p(Y^{(b)}|x_k^{(b)} = -1)P(x_k^{(b)} = -1)}. \tag{3.5}$$

Define the set $\mathcal{Z}_{k,+1}$ to be the set of symbols $S^{(b)}$, where the k^{th} bit in the label $\mathbf{x}^{(b)}$ of these set of symbols is $+1$. That is,

$$\mathcal{Z}_{k,\pm 1} = \{S^{(b)}|x_k = \left[\mathcal{M}^{-1}(S^{(b)})\right]_k = \pm 1, S^{(b)} \in \mathcal{C}\}, \tag{3.6}$$

where $[\cdot]_k$ denotes the k^{th} element of the vector. Using (3.6), the *a posteriori* soft information $L_D(x_k^{(b)}|Y^{(b)})$ can be expressed as

$$L_D(x_k^{(b)}|Y^{(b)}) = \ln \frac{\sum_{S^{(b)} \in \mathcal{Z}_{k,=+1}} p(Y^{(b)}|S^{(b)})P(S^{(b)})}{\sum_{S^{(b)} \in \mathcal{Z}_{k,=-1}} p(Y^{(b)}|S^{(b)})P(S^{(b)})}, \tag{3.7}$$

where $p(Y^{(b)}|S^{(b)})$ is the likelihood of the received signal given a transmitted matrix. If the interleaver in the BICM-IDD scheme is well designed and the block length of the outer code $\mathbf{x}$ is long (typical lengths range from 2000 [22] to 10000 [16]), the

n bits in a given block $\mathbf{x}^{(b)}$ are approximately independent. Thus, $P(S^{(b)})$ can be approximated by

$$P(S^{(b)}) \approx \prod_{k=1}^{n} P(x_k = [\mathcal{M}^{-1}(S^{(b)})]_k). \tag{3.8}$$

In the first iteration of the iterative demapping and decoding scheme, it is assumed that the bits are equiprobable and hence the $P(S^{(b)}) = 1/2^n$; i.e., the *a priori* information L_{A_1} in the first iteration is set to be the all zero vector. In the subsequent iterations, the *a priori* soft information L_{A_1} is the interleaved extrinsic soft information generated by the outer decoder in the previous iteration; namely L_{E_2}. In the next section, we will illustrate the principles of BICM-IDD for a coherent MIMO system with V-BLAST signalling.

3.3 Example: BICM-IDD for a Coherent MIMO System with V-BLAST Signalling

In this section, we will describe a simple technique to communicate over a MIMO channel, namely V-BLAST (Vertical Bell Laboratories Layered Space-Time Architecture). V-BLAST systems attempt to exploit the multiple paths inherent in the MIMO system in order to operate at high data rates. The V-BLAST system transmits a vector of symbols given by the $M \times 1$ vector $\mathbf{s}$ whose entries are chosen from an integer constellation $\mathcal{C}$ (e.g. PSK, QAM). The V-BLAST system assumes that the receiver has access to perfect CSI and that the time dimension of the space-time block code is equal to 1. We consider a system with M transmit antennas and N receive antennas over a richly-scattered flat fading channel, where $N \geq M$. The received channel symbols $N \times 1$ vector $\mathbf{y}$ can be written as

$$\mathbf{y} = H\mathbf{s} + \sqrt{\frac{M}{\rho}}\mathbf{v}, \tag{3.9}$$

where H is the $N \times M$ channel matrix, $\mathbf{v}$ is the $N \times 1$ vector representing the additive noise whose entries are drawn independently from the standard complex Gaussian distribution $\mathcal{CN}(0,1)$ and ρ is the signal to noise ratio.

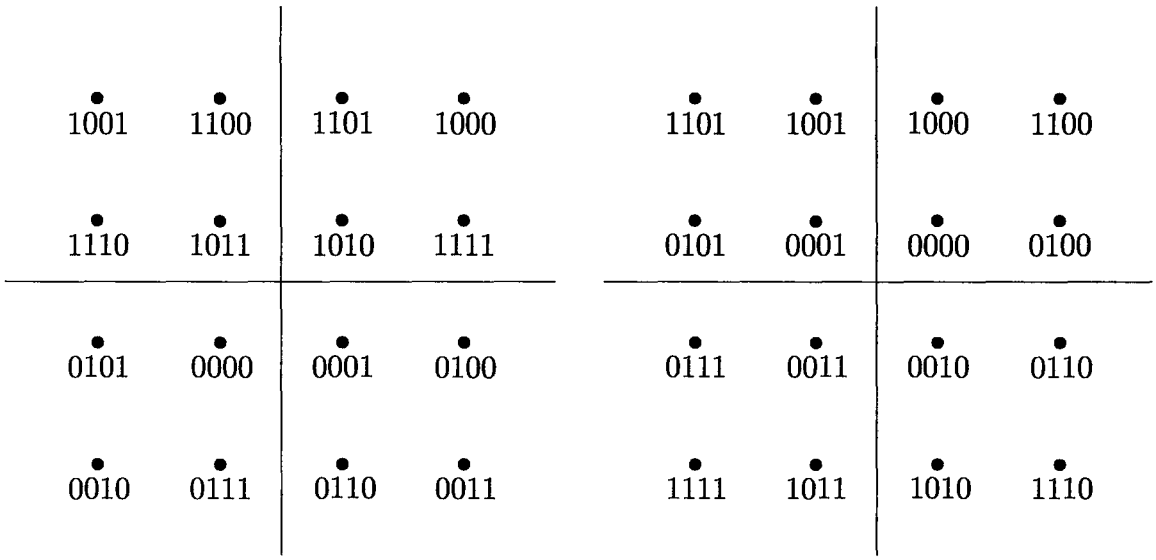
In the constellation mapper, the n-bit vector $\mathbf{x}^{(b)}$ that is to be mapped to a space-time symbol is divided into M sequences of ℓ bits $\mathbf{x}_i^{(b)}$, where each $\mathbf{x}_i^{(b)}$ will be mapped onto a scalar symbol $s_i^{(b)}$ from a standard integer constellation. The vector of transmitted symbols that corresponds to the b^{th} block of $M\ell$ bits in $\mathbf{x}$ can be written as

$$\mathbf{s}^{(b)} = \mathcal{M}(\mathbf{x}^{(b)}) = [s_1^{(b)}, s_2^{(b)}, ..., s_M^{(b)}]^T = [\mathcal{M}_1(\mathbf{x}_1^{(b)}), \mathcal{M}_2(\mathbf{x}_2^{(b)}), ..., \mathcal{M}_M(\mathbf{x}_M^{(b)})]^T, \quad (3.10)$$

where $(\cdot)^T$ is the transpose of the vector and $\mathcal{M}_i(\mathbf{x}_i^{(b)})$ denotes the mapping of the ℓ-bit $\mathbf{x}_i^{(b)}$ sequence onto a single point on the constellation $\mathcal{C}$.

There are different techniques to map a sequence of bits onto a constellation. Two well known techniques are set partitioning and Gray labeling. Set partitioning was introduced in [45] for trellis coded modulation. Set partitioning assigns bits based on the successive partitioning of the constellation into subsets with increasing minimum Euclidean distance [14]. An example of a 16-QAM set-partitioned constellation is shown in Figure 3.5(a). An alternative mapping technique is Gray labeling. Gray labeling maps the bits onto the constellation $\mathcal{C}$ such that for each constellation point which lies at the minimum Euclidean distance from any point on the constellation, its label $\mathbf{x}_i^{(b)}$ differs by only one bit from the label of this point. This can be easily observed from Figure 3.5(b), which is a Gray labeled 16-QAM constellation.

The vector of constellation symbols $\mathbf{s}^{(b)}$ indexed by the interleaved coded bits in the block $\mathbf{x}^{(b)}$ are then transmitted through a multiple antenna channel (3.9), which transmits M constellation symbols for each channel use. In the BICM scheme, if the outer code rate is R then the communication data rate is given by $RM\ell$ bits per channel use.



(a) Set partitioning (b) Gray labeling

Figure 3.5: 16-QAM constellation (a) Set partitioning and (b) Gray labeling, [14].

At the receiver, the iterative demapping and decoding scheme described in Section 3.2.2 is applied. Assuming the channel model in (3.9), the *a posteriori* soft information in (3.7) can be expressed as

$$L_D(x_k^{(b)}|\mathbf{y}^{(b)}) = \ln \frac{\sum_{\mathbf{s}^{(b)} \in \mathcal{Z}_{k,=+1}} p(\mathbf{y}^{(b)}|\mathbf{s}^{(b)}) P(\mathbf{s}^{(b)})}{\sum_{\mathbf{s}^{(b)} \in \mathcal{Z}_{k,=-1}} p(\mathbf{y}^{(b)}|\mathbf{s}^{(b)}) P(\mathbf{s}^{(b)})}, \tag{3.11}$$

where the likelihood of the received signal given a transmitted matrix is given by

$$p(\mathbf{y}^{(b)}|\mathbf{s}^{(b)}) = \frac{\exp(-\frac{\rho}{2M}\|\mathbf{y}^{(b)} - H\mathbf{s}^{(b)}\|^2)}{(2\pi M/\rho)^N}, \tag{3.12}$$

and the *a priori* soft information from the decoder can be approximated by

$$P(\mathbf{s}^{(b)}) \approx \prod_{k=1}^{M\ell} P(x_k = [\mathcal{M}_1^{-1}(s_1^{(b)}), \mathcal{M}_2^{-1}(s_2^{(b)}), ..., \mathcal{M}_M^{-1}(s_M^{(b)})]_k). \tag{3.13}$$

As can be seen from (3.11), the cost of computing the APP for each bit grows exponentially in the number of bits per block. The number of bits per block is

the product of the number of symbols per block, M, and the number of bits per symbol, ℓ. This can be a severe computational burden on the demapper and hence exact MAP demapping is computationally infeasible, especially for systems with a large constellation size [16]. To alleviate the complexity burden that is generated by performing an exhaustive search on all the combinations of symbols of a constellation, one may choose to perform such computations on only a subset of those symbols. Such a technique was suggested in [16], where a list sphere demapper generates a list of constellation symbol blocks of predetermined size, N_{cand}, for each transmitted block of symbols $\mathbf{s}^{(b)}$. There is a tradeoff between the size of the list and the reliability of the detection. A larger list results in a more reliable system, but also incurs increased computational complexity. It was suggested in [16] that the number of candidate symbol blocks N_{cand} in the list should be as large as possible while having acceptable complexity. This list-based technique will provide some of the motivation for the development of a list-based demapper for the non-coherent MIMO system in Section 5.3.1.

In this chapter, the iterative demapping and decoding scheme for BICM was described and particular attention was given to the MIMO demapper. The next chapter will focus on the outer error control code and its soft in/soft out decoding algorithm which will be used to generate the *a posteriori* soft information L_{D_2} in the BICM-IDD scheme.

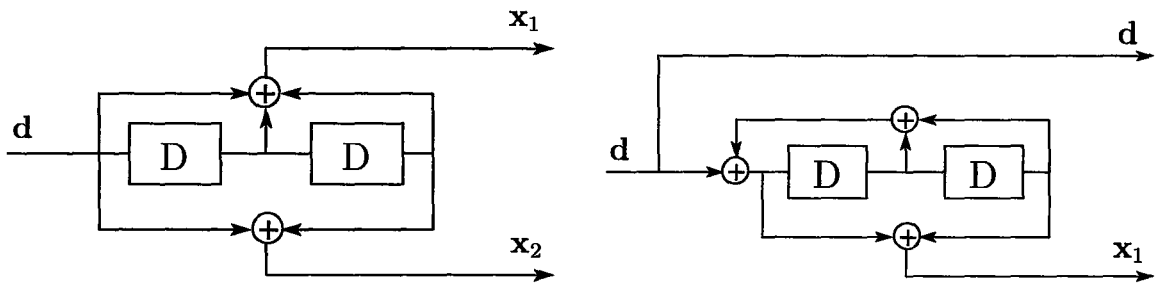
Chapter 4

Turbo Codes

4.1 Introduction

In the previous chapter, the principles of the BICM-IDD scheme were described. The aim of this scheme is to offer reliable communication at data rates close to the ergodic capacity of the channel [15, 16]. Within the building blocks of this scheme, there is an outer error-control encoder and its corresponding soft decoder that allow the BICM-IDD scheme to communicate reliably near the capacity limit. These error-control codes introduce a structure between the transmitted bits by adding additional redundancy bits to the source bits. The channel encoder constrains the bit sequence so that only a strict subset of all possible sequences can be transmitted. This structure can then be exploited by the receiver to improve the performance of the communication system by mitigating the effects of noise and channel variations.

A class of highly effective codes can be generated by concatenating simple convolutional codes. Such codes are referred to as turbo codes [38]. The power of turbo coding lies in its low-complexity "turbo-like" decoding technique, in which each constituent concatenated code is decoded separately and soft information is exchanged iteratively between the respective decoders. The BICM-IDD scheme described in



(a) Non-systematic code (b) Recursive systematic code

Figure 4.1: Rate 1/2 convolutional codes: (a) Non-systematic (b) Recursive systematic.

Section 3.2.2 also uses this "turbo" technique to iteratively exchange soft information between the demapper and the decoder.

The purpose of this chapter is to provide a detailed description of turbo codes and their building blocks. Turbo codes can be implemented through the parallel concatenation of recursive systematic constituent convolutional encoders and we will focus on this case. Section 4.2 will describe recursive systematic convolutional codes. Section 4.3 will describe the structure of the turbo encoder, while Section 4.4 will discuss the iterative soft-input/soft-output decoding technique for turbo codes. Section 4.4.2 will describe the computation of the *a posteriori* soft information L_{D_2} in the BICM-IDD scheme.

4.2 Convolutional Codes

A convolutional encoder is a finite memory shift register system in which a continuous stream of data bits are input into its shift register and a stream of output coded bits is generated. The coded bits contain additional redundancy bits that act as parity check bits at the receiver. Convolutional encoders can be categorized as being either systematic or non-systematic, a systematic encoder is an encoder in which the input

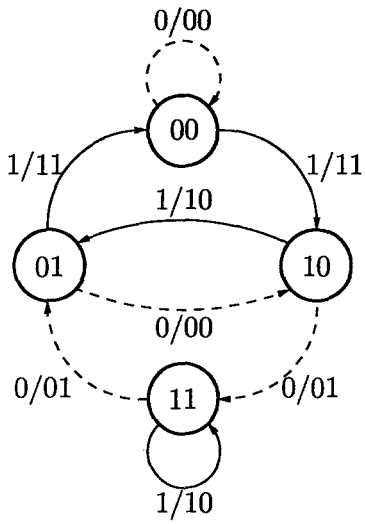


Figure 4.2: State transition diagram of the rate 1/2 recursive systematic encoder shown in Figure 4.1(b). The dashed line represents an input bit"0" while the solid line represents an input bit "1".

data bits appear unchanged in the coded output bits; see Figure 4.1. Convolutional encoders can also be categorized as being either recursive or non-recursive, a recursive encoder is an encoder where at least one feedback loop appears between the blocks of the shift register, as shown in Figure 4.1.

The content of the shift register of a convolutional encoder at a given time is called the current state s. Each convolutional code can be represented by a state transition diagram that displays all the the transitions between the states based on the input of the encoder. The state transition diagram of the rate 1/2 recursive systematic convolutional code in Figure 4.1(b) is shown in Figure 4.2. The state transitions are labeled with the (input/output) pairs of the encoder. An alternative beneficial representation is the trellis diagram, which represents all the possible state progressions as time passes; see Figure 4.3. Each path on the trellis represents a unique input sequence $[d_1, d_2, ..., d_L]$, where L is the length of the input data block. Trellis diagrams usually start from the all zero state and end at the same state. In order to terminate the trellis of a non-recursive convolutional code at the all zero

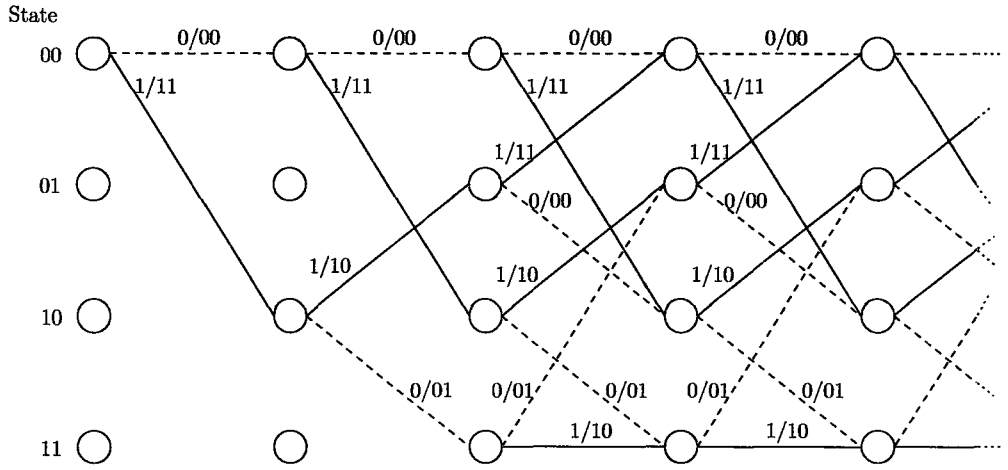


Figure 4.3: Trellis diagram of the rate 1/2 recursive systematic encoder shown in Figure 4.1(b).

state, a block of zeros are inserted into the shift register after the data block. The length of this block is equal to the number of memory blocks ν in the shift register. However, this is not the case with recursive convolutional codes which can be rather difficult to terminate at the all zero state.

4.3 Turbo Encoder

The first turbo encoder introduced in [38] consisted of two recursive systematic binary convolutional codes concatenated in parallel, separated by an interleaver that scrambles the bits in a pseudo-random fashion. An example of a rate 1/3 recursive systematic turbo code is illustrated in Figure 4.4, where the two rate 1/2 constituent codes are shown in Figure 4.1(b). In this scheme, the input data bits are passed into the first constituent code, which generates a systematic codeword in the form $[\mathbf{d}, \mathbf{x}_1]$, where $\mathbf{d}$ represents the input data bits, and $\mathbf{x}_1$ are the parity check bits from the first constituent convolutional code. The input data bits are then interleaved and passed into the second constituent convolutional code, which generates the second

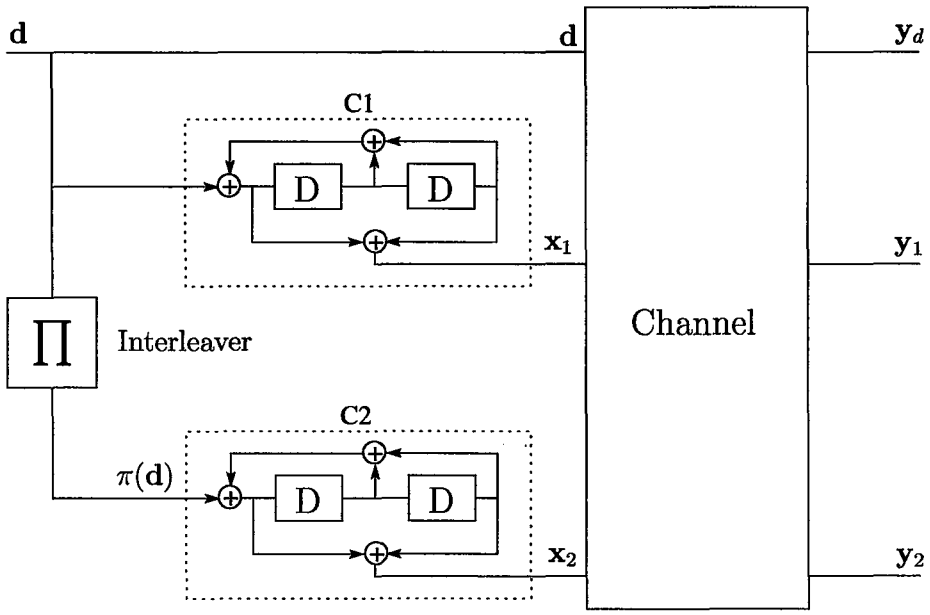


Figure 4.4: A rate 1/3 recursive systematic turbo code.

set of parity check bits, $\mathbf{x}_2$. Thus the output of the turbo encoder can be written in the form of $[\mathbf{d}, \mathbf{x}_1, \mathbf{x}_2]$.

The choice of the interleaver in the turbo encoder has a strong influence on the performance of the turbo coded system, particularly when the block length is short, for example of length less than 1000, [46]. The commonly used pseudo-random interleaver can perform poorly for such short blocks. Several deterministic approaches have been proposed in [46–48] that generate interleavers that significantly outperform pseudo-random interleavers. For longer block lengths, it was shown in [48] that pseudo-random interleavers provide excellent performance with high probability. In fact, it was shown that deterministic approaches in [46–48] perform only as well as the average performance of a set of pseudo-random interleavers. In this thesis, a long block length of 8000 data bits was considered. Therefore for simplicity, a set of 10 (different) pseudo-random interleavers were examined via simulations, and the interleaver with the best performance was chosen.

Since it is usually desirable to increase the data rate of communication systems, one way is by increasing the data rate of the turbo encoder. However, increasing the data rate of turbo codes decreases the number of parity check bits which will often result in a weaker structure between the encoded bits. A common technique to increase the rate of the turbo code is through puncturing. Puncturing discards a fraction of the parity check bits according to a certain pattern; e.g., the rate 1/3 turbo code shown in Figure 4.4 can be punctured by alternately discarding the parity check bits generated by one of the constituent codes for each data bit to produce a rate 1/2 turbo code. In the next section, the turbo decoding technique will be described.

4.4 Decoding Turbo Codes

In order to operate at rates close to the capacity limit using turbo codes, the data blocks are usually rather long (typical lengths range from 3500 [20] to 65536 [38]). Applying joint optimum decoding to turbo codes is prohibitively complex [4]. As an alternative, the turbo decoder attempts to approximate the optimum decoder by decoding each constituent code separately while iteratively exchanging soft information between the decoding blocks; see Figure 4.5.

As shown in Figure 4.5, the decoder of the first constituent code, DEC1, uses the channel observations of the systematic component and the output of the first constituent code $[\mathbf{y_d}, \mathbf{y_1}]$ and *a priori* information from the decoder of the second constituent code, DEC2, namely L_{P_1}, to compute the *a posteriori* soft information about each bit given the received observation L_{U_1}. Similarly, DEC2 uses the *a priori* soft information from DEC1, namely L_{P_2}, and the channel observations of the interleaved systematic component and the output of the second constituent code, namely $[\pi(\mathbf{y_d}), \mathbf{y_2}]$, to compute the *a posteriori* soft information L_{U_2}.

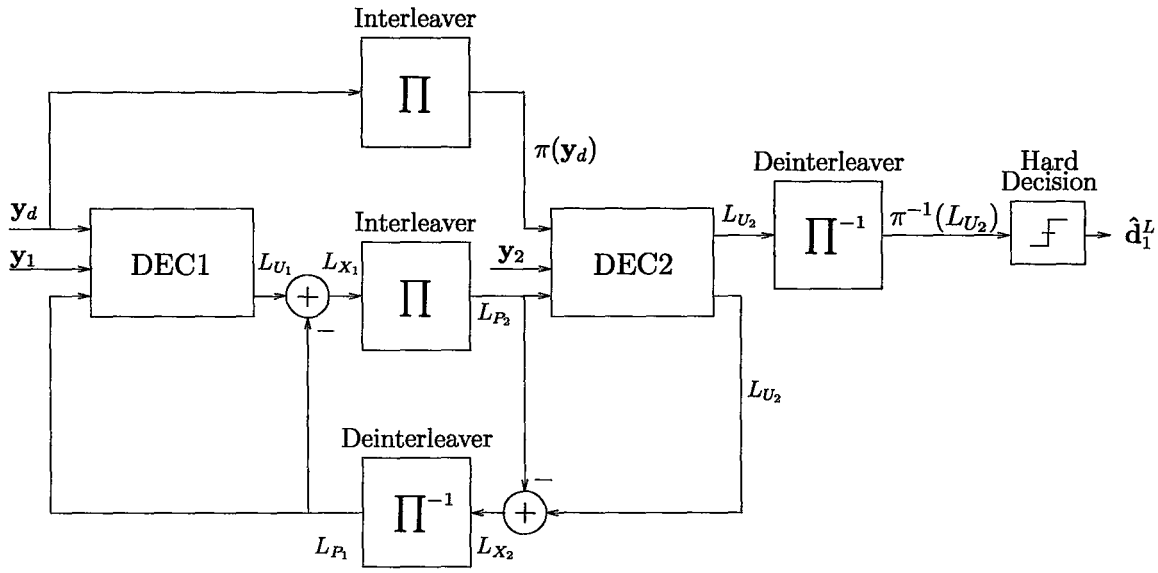


Figure 4.5: The turbo decoding algorithm.

Within each decoding block, a soft input/soft output algorithm is needed to decode the received signal by exploiting the structure introduced by the corresponding component of the encoder. A well known technique that minimizes the bit error probability of each constituent code is the BCJR algorithm [39]. The BCJR algorithm computes the *a posteriori* probability (APP) for each bit using the structure of the constituent code. This soft APP information generated by the BCJR algorithm is then iteratively exchanged between the constituent decoding blocks to provide an efficient decoding technique for turbo codes.

4.4.1 The BCJR Algorithm

The BCJR algorithm was developed by Bahl et al. [39] to provide a decoding technique for convolutional codes that minimizes the bit error probability by maximizing the APP for each bit. A simple way to describe this algorithm is by looking at the trellis of the constituent code, which consists of branches that represents the allowable state transitions provided by the code as time progress. Let us consider the rate 1/2

recursive systematic convolutional code shown in Figure 4.1(b), and its trellis in Figure 4.3. The input bits to the constituent encoder are denoted by $\mathbf{d}_1^L = [d_1, d_2, ..., d_L]$, where L is the length of the input block. Each input bit d_t is associated with a transition between the states S_{t-1} and S_t, where S_t is the encoder state at time t. The total number of states is 2^ν, where ν represents the code memory. The encoder's output symbol for each input bit d_t is denoted by $\mathbf{x}_t = (x_t^{(1)}, x_t^{(2)})$, where $t = 1, 2, ..., L$, and since the convolution encoder is systematic $x_t^{(1)} = d_t$. Let us denote the channel observation symbol by $\mathbf{y}_1^L = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_L]$, where $\mathbf{y}_t = (y_t^{(1)}, y_t^{(2)})$.

The APP for each bit can be easily computed if the *a posteriori* transition probability for all state transitions in the trellis are known. The *a posteriori* transition probability between states s' and s at time t is given by [39]

$$P(S_{t-1} = s', S_t = s|\mathbf{y}_1^L) = \frac{p(S_{t-1} = s', S_t = s, \mathbf{y}_1^L)}{p(\mathbf{y}_1^L)}. \tag{4.1}$$

Since convolutional codes with i.i.d inputs can be viewed as a discrete-time finite-state Markov chain [4], if S_t is known, events occurring after time t do not depend on $\mathbf{y}_1^t$ [39]. The numerator in (4.1) can therefore be written as,

$$\begin{aligned} \sigma_t(s', s) &= p(S_{t-1} = s', \mathbf{y}_1^{t-1}) p(S_t = s, \mathbf{y}_t|S_{t-1} = s') p(\mathbf{y}_{t+1}^L|S_t = s) \\ &= \alpha_{t-1}(s')\gamma_t(s', s)\beta_t(s), \end{aligned} \tag{4.2}$$

where $\alpha_t(s) = p(S_t = s, \mathbf{y}_1^t)$, $\beta_t(s) = p(\mathbf{y}_{t+1}^L|S_t = s)$ and $\gamma_t(s', s) = p(S_t = s, \mathbf{y}_t|S_{t-1} = s')$. As shown in equation (4.2), the BCJR algorithm decomposes the *a posteriori* probability into three factors: $\alpha_{t-1}(s')$ which depends on the past observations; $\beta_t(s)$ which depends on the future observations; and $\gamma_t(s', s)$ which depends on the present observation; see Figure 4.6.

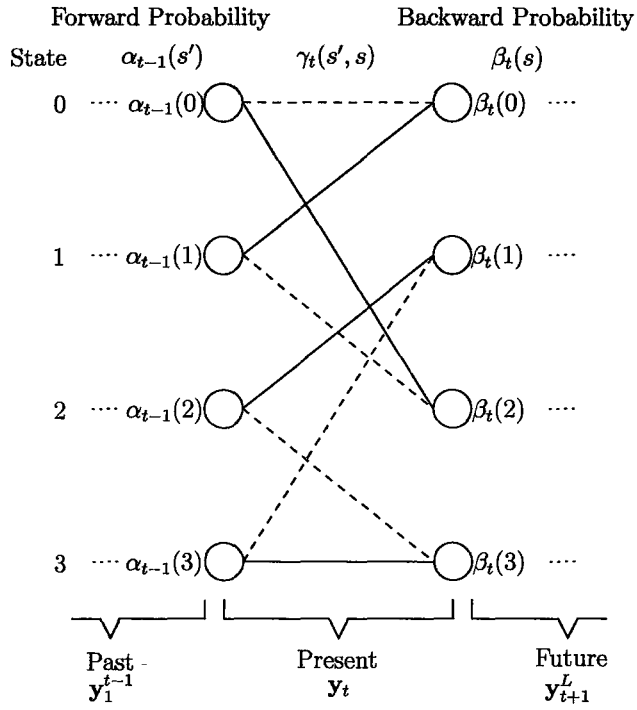


Figure 4.6: BCJR algorithm examines the trellis transitions between states at time t.

The probability $\alpha_t(s)$ can be computed recursively as shown in [39],

$$\begin{aligned}
\alpha_t(s) &= \sum_{s'=0}^{2^\nu-1} p(S_{t-1}=s', S_t=s, \mathbf{y}_1^t) \\
&= \sum_{s'=0}^{2^\nu-1} p(S_{t-1}=s', \mathbf{y}_1^{t-1}) p(S_t=s, \mathbf{y}_t | S_{t-1}=s') \\
&= \sum_{s'=0}^{2^\nu-1} \alpha_{t-1}(s')\gamma_t(s',s). \qquad (4.3)
\end{aligned}$$

Similarly, the probability $\beta_t(s)$ can also be computed in a recursive fashion,

$$\begin{aligned}
\beta_t(s) &= \sum_{s'=0}^{2^\nu-1} p(S_{t+1}=s', \mathbf{y}_{t+1}^L | S_t=s), \\
&= \sum_{s'=0}^{2^\nu-1} p(S_{t+1}=s', \mathbf{y}_{t+1} | S_t=s) p(\mathbf{y}_{t+2}^L | S_{t+1}=s') \\
&= \sum_{s'=0}^{2^\nu-1} \beta_{t+1}(s')\gamma_{t+1}(s,s'). \qquad (4.4)
\end{aligned}$$

The recursive fashion in which the probabilities $\alpha_t(s)$ and $\beta_t(s)$ are generated can be described by a two-way propagation through the trellis, a forward propagation to compute $\alpha_t(s)$ and a backward propagation to compute $\beta_t(s)$. Since convolutional codes start from the all zero state and usually terminate at the all zero state, in this case the boundary conditions for the generation of $\alpha_t(s)$ and $\beta_t(s)$ can be represented by [39]

$$\alpha_0(0) = 1, \qquad \alpha_0(s) = 0, \quad \forall s = 1, ..., 2^\nu - 1, \tag{4.5}$$

$$\beta_{L+\nu}(0) = 1, \qquad \beta_{L+\nu}(s) = 1, \quad \forall s = 1, ..., 2^\nu - 1, \tag{4.6}$$

Unlike non-recursive convolutional codes, for which padding ν zeros is sufficient to put the encoder in the all zero state, it can be rather difficult to put the encoder of a recursive code to the all zero state at a given time. An alternative initialization of the backward boundary condition of "unterminated" recursive convolutional codes was suggested in [38]. This boundary condition is based on the fact that the trellis can terminate in any state with equal probability since the data input bits are independent and equally probably. Therefore, the backward boundary condition can be given by [38],

$$\beta_L(s) = \frac{1}{L}, \quad \forall s = 0, ..., 2^\nu - 1. \tag{4.7}$$

Each branch in the trellis is associated with a certain probability $\gamma_t(s', s)$ that describes the the transition probability between states s' and s at a given time t. This can be written as

$$\gamma_t(s', s) = p(\mathbf{y}_t | S_{t-1} = s', S_t = s) P(S_t = s | S_{t-1} = s'). \tag{4.8}$$

The probability $p(\mathbf{y}_t | S_{t-1} = s', S_t = s)$ is the transition probability associated with the channel. Assuming that the transition between s' and s exists, the probability $P(S_t = s | S_{t-1} = s')$ describes the probability of the encoder input bits. Generally, the input bits of the encoder are equally likely, i.e., $P(d_t = 0) = P(d_t = 1) = 1/2$.

However, in turbo decoding, where *a priori* information is exchanged between the BCJR decoding blocks, it will be shown in Section 4.4.2 that $P(S_t = s|S_{t-1} = s')$ represents the *a priori* information from the other constituent BCJR decoder.

In order to minimize the bit error probability, the decoder has to find the most likely coded bit $x_t^{(i)}$ given the received sequence $\mathbf{y}_1^L$. Using (4.1) and (4.2), the APP of each bit is given by [39],

$$P(x_t^{(i)} = 0|\mathbf{y}_1^L) = \frac{\sum_{(s',s)\in\mathcal{Q}_{t,0}^{(i)}} \sigma_t(s',s)}{p(\mathbf{y}_1^L)}, \tag{4.9}$$

$$= \frac{\sum_{(s',s)\in\mathcal{Q}_{t,0}^{(i)}} \alpha_{t-1}(s')\gamma_t(s',s)\beta_t(s)}{p(\mathbf{y}_1^L)}, \tag{4.10}$$

where $\mathcal{Q}_{t,0}^{(i)}$ is the set of transitions between $S_{t-1} = s'$ and $S_t = s$ such that the i^{th} encoder output bit $x_t^{(i)}$ that labels any transition in the set is equal to 0.

In order to complete the decoding operation of a convolutional code, a hard decision would be performed on the APP probability. That is, if $P(x_t^{(i)} = 0|\mathbf{y}_1^L)$ is greater that 0.5 then $x_t^{(i)} = 0$, otherwise $x_t^{(i)} = 1$. However, as mentioned in Section 4.4, in order to decode turbo codes, the constituent decoders exchange soft information in an iterative fashion to approximate the optimal joint decoder. Therefore, this hard decision is performed only after a number of iterations between the constituent decoders. In the next section, we will describe the turbo decoding process using the BCJR algorithm.

4.4.2 Turbo Decoding using the BCJR algorithm

In order to perform turbo decoding, soft information is iteratively exchanged between the constituent decoding blocks, as shown in Figure 4.5. Thus, as discussed in Section 3.2.2, it is convenient to express this soft information in the form of log likelihood ratios (L-values) [44]. Using (4.9), the L-value representing the APP for each coded

bit generated by one constituent BCJR algorithm can be written as [38],

$$L_C(x_t^{(i)}|\mathbf{y}_1^L) = \ln \frac{P(x_t^{(i)}=1|\mathbf{y}_1^L)}{P(x_t^{(i)}=0|\mathbf{y}_1^L)} \tag{4.11}$$

$$= \ln \frac{\sum_{(s',s)\in\Omega_{t,1}^{(i)}} \sigma_t(s',s)}{\sum_{(s',s)\in\Omega_{t,0}^{(i)}} \sigma_t(s',s)} \tag{4.12}$$

$$= \ln \frac{\sum_{(s',s)\in\Omega_{t,1}^{(i)}} \alpha_{t-1}(s')\gamma_t(s',s)\beta_t(s)}{\sum_{(s',s)\in\Omega_{t,0}^{(i)}} \alpha_{t-1}(s')\gamma_t(s',s)\beta_t(s)}. \tag{4.13}$$

In order to compute equation (4.11), the BCJR uses the channel observations and the *a priori* information from the other constituent decoder to compute the branch transition probability $\gamma_t(s',s)$ in (4.11). In the first iteration, the bits are assumed to be equally likely; i.e., the *a priori* soft information L_{P_1} is set to equal the all zero vector. Due to the structure of the turbo encoder in Figure 4.4, only a subset of the APP soft information in L_C is beneficial to the other decoder; namely the APP soft information for the information bits in the code. Since the information bits of a systematic code appear unchanged in the sequence of output coded bits, the APP soft information of the systematic bits consists of the entries in L_C that corresponds to those bits.

The iterative exchange of soft information shown in Figure 4.5 can be described as follows. After Decoder1 (DEC1) computes the APP for the coded bits L_{C_1} using the BCJR algorithm as shown in (4.11), DEC1 selects the entries that corresponds to the systematic bits L_{U_1}. The *a priori* soft information L_{P_1} is subtracted from the systematic APP soft information L_{U_1}, generating the extrinsic soft information L_{X_1}. The extrinsic information L_{X_1} is interleaved generating the *a priori* soft information L_{P_2} for Decoder2 (DEC2). Similarly, DEC2 generates the extrinsic soft information for the systematic bits L_{X_2}, which is deinterleaved to become the *a priori* soft information L_{P_1} for DEC1; completing an iteration. The extrinsic uncoded soft information is then exchanged iteratively between the constituent decoders to improve

the performance of the decoding process. After a certain number of iterations, a hard decision is performed on the deinterleaved L-value representing the uncoded APP soft information of DEC2, namely $\pi^{-1}(L_{U_2})$, to obtain the decoded output $\hat{\mathbf{d}}_1^L$ as shown in Figure 4.5.

However, in the BICM-IDD scheme described in Chapter 3 and shown in Figure 3.3, soft information is also exchanged between the demapper and the decoder to improve the performance of the detection process. The iterations between the demapper and the decoder can be viewed as the outer iteration while the iterations between the constituent decoders in turbo decoding can be viewed as the inner iterations. That is, during each demapper to decoder iteration, the constituent decoders exchange soft information for several iterations. Unlike turbo decoding in which only the systematic APP soft information is exchanged between the decoding blocks, the exchange of soft information between the demapper and the decoder in the BICM-IDD scheme requires the APP soft information of the parity bits as well. In other words, the APP soft information $\pi^{-1}(L_{U_2})$ which represents the systematic bits are combined with entries from the APP soft information L_{C_1} and L_{C_2} that corresponds to the coded bits $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively to obtain the APP soft information L_{D_2} for all bits. That is, L_{D_2} is of the form $\mathring{\text{o}}f[\pi^{-1}(L_{U_2}), L_{C_{1,j}}, L_{C_{2,k}}]$, where j and k represent the indices of the entries in L_{C_1} and L_{C_2} that corresponds to the bit positions of the coded bits $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. The decoder's APP soft information L_{D_2} is then exchanged iteratively with the demapper's APP soft information L_{D_1} as described in Section 3.2.2; see Figure 3.3.

This chapter provided an overview of turbo codes and their BCJR soft iterative decoding technique, which can be used as the outer phase of the BICM-IDD scheme. In the next chapter, we will use the principles of BICM-IDD and non-coherent MIMO systems to build a high performance non-coherent MIMO communication scheme.

Chapter 5

BICM Scheme for Non-Coherent MIMO Communication Systems

5.1 Introduction

The focus of this thesis is to develop a computationally-efficient communication scheme that, at high SNRs, provides low error rates at data rates close to the ergodic capacity of the non-coherent MIMO channel. One way to construct such a scheme is by applying the principles of BICM with iterative demapping and decoding (IDD) to the non-coherent multiple antenna channel. In the previous chapters, the building blocks of such a system were addressed. Chapter 2 discussed the properties of non-coherent multiple antenna channels. Chapter 3 described the principles of the BICM-IDD scheme, while Chapter 4 described turbo coding and the corresponding iterative soft decoder that can be used as the outer error-control code in the BICM-IDD scheme.

In this chapter, these principles will be combined to construct a BICM-IDD scheme for the non-coherent MIMO channel. A block diagram of the proposed scheme is shown in Figure 5.1. We will discuss the challenges that were encountered in the

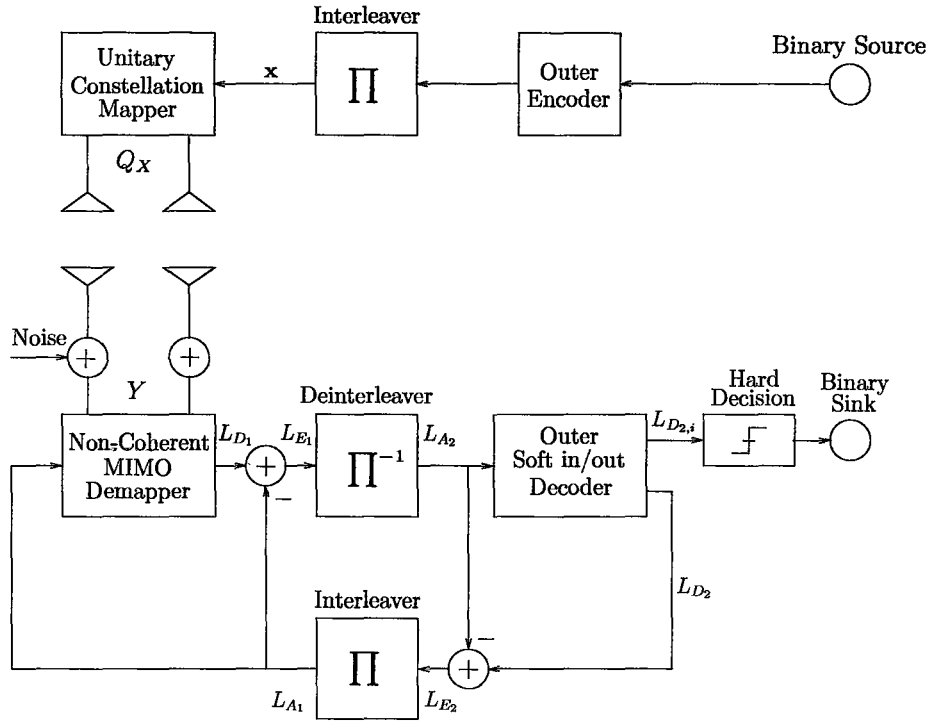


Figure 5.1: The BICM-IDD scheme for the non-coherent MIMO channel.

construction of this scheme and we will describe the methods by which they were overcome. The main contributions of this work lie in the non-coherent constellation mapper and the corresponding non-coherent MIMO demapper. The next section, will describe the unitary constellation mapper. Section 5.3 will describe the iterative demapping and decoding technique for the non-coherent BICM-IDD scheme, which will give rise to a computationally efficient list-based demapper. In addition, a technique to incorporate the decoder's information into the demapper's list construction technique will be proposed, and its significant improvement in performance will be demonstrated. In Section 5.4, several parameters that affect the performance of the BICM-IDD scheme will be addressed. Finally, in Section 5.5 the proposed scheme will be compared to an alternative training-based scheme for the non-coherent MIMO channel.

5.2 Non-coherent MIMO Mapper

The non-coherent MIMO mapper maps consecutive blocks of n bits from the block of interleaved coded bits $\mathbf{x}$ onto a space-time constellation point. As described in Section 2.1, at high SNRs the non-coherent capacity achieving channel symbols are in the form of isotopically distributed unitary matrices Q_X. The information is conveyed in the subspace spanned by each matrix Q_X and each subspace can be represented as a "constellation" point on a compact Grassmann manifold [12]. Before being able to map the interleaved coded bits $\mathbf{x}$ onto constellation points, the first challenge is to design a constellation that mimics the capacity achieving distribution. As discussed in Section 2.3, a constellation of isotropically distributed unitary matrices Q_X can be directly generated by maximizing the mutual distances between the subspaces spanned by all the matrices Q_X in the constellation. Unfortunately, the complexity of this direct approach increases rapidly as the size of the constellation grows. An alternative greedy technique was developed [27]. This greedy algorithm recursively adds one constellation matrix Q_X that essentially maximizes the minimum of the distances to all the subspaces spanned by the matrices that are already in the constellation.

After generating a constellation of size 2^n using the greedy algorithm, the second challenge lies in finding a technique to map consecutive blocks of n (interleaved) coded bits to points in the Grassmannian constellation. If we let $\mathbf{x}_b$ denote the b^{th} length-n block of $\mathbf{x}$ in Figure 5.1, the mapping will be denoted by $Q_X^{(b)} = \mathcal{M}(\mathbf{x}_b)$, where $Q_X^{(b)}$ is the b^{th} transmitted unitary matrix and $\mathcal{M}(\cdot)$ is the mapping function for a length-n vector of bits onto a unitary matrix in the constellation.

In the case of coherent MIMO communication systems, in which a standard constellation is usually used, several numerically optimized mappings have recently been proposed [23, 49, 50]. However, due to the Grassmannian geometry of the proposed non-coherent MIMO scheme, and the imperfections in the constellation generated by

the greedy algorithm, there is no conventional technique to label the points in the Grassmannian constellation. In [23] a labelling technique was proposed in which a cost function based on the Chernoff bound of the pairwise error probability is minimized. The numerically optimized mapping performs well in the case of differential schemes, in which the channel variations between two consecutive blocks can be considered negligible. However, in the considered block-fading model this optimized mapping does not show any improvement in performance over pseudo-random mapping [23].

In this thesis, a simple mapping technique that incorporates the structure of the Grassmannian constellation is proposed. It was shown in Section 2.3 that for the case where the coherence time is equal to twice the number of transmitters (i.e., $T = 2M$), the Grassmannian constellation consists of pairs of maximally separated points; i.e. the maximum distance is between the subspace spanned by each Q_X and its orthogonal complement $Q_X^{\perp}$. Therefore, the proposed mapping technique will partition the constellation so that constellation pairs that are separated by the maximum chordal Frobenius distance have the minimum Hamming distance in their label; i.e., their labels have Hamming distance 1. We partition the 2^n point constellation into two subsets of size 2^{n-1} with one element of each maximally separated pair in each subset. The first $n-1$ bits in $\mathbf{x}_b$ are mapped pseudo-randomly to the 2^{n-1} maximally seperated constellation pairs and then the remaining bit in $\mathbf{x}_b$ is used to select one element of the chosen pair (a certain subset). In this thesis, a set of 10 (different) pseudo-random mappings were examined via simulations, and the mapping with the best performance was chosen. The geometry of the imperfect Grassmannian lattices generated by the greedy constellation design technique in Section 2.3 has restricted the sophistication of the proposed mapping technique, but the simulation results in Section 5.5 suggest that in spite of its simplicity, the proposed mapper provides good performance.

5.3 Iterative Demapping and Decoding for Non-Coherent Systems

As discussed in Section 3.2.2, the iterative demapping and decoding process in a BICM-IDD scheme is a sequential process in which the "extrinsic" soft information from the previous decoder iteration is used by the demapper to extract soft information from each channel use. The "extrinsic" component of this information is subsequently passed to the decoder for its next iteration. If we let $x_k^{(b)}$ denote the k^{th} element of $\mathbf{x}_b$, the b^{th} length-n block of $\mathbf{x}$ in Figure 5.1, then the demapper's role is to compute the *a posteriori* L-value of $x_k^{(b)}$ conditioned on the received matrix $Y^{(b)}$ for $k = 1, 2, ..., n$. That is, the demapper computes

$$L_{D_1}(x_k^{(b)}|Y^{(b)}) = \ln \frac{P(x_k^{(b)} = +1|Y^{(b)})}{P(x_k^{(b)} = -1|Y^{(b)})}. \tag{5.1}$$

The computation of the L-value in (5.1) can be simplified by using Bayes rule,

$$L_{D_1}(x_k^{(b)}|Y^{(b)}) = \ln \frac{\sum_{Q_X \in \mathcal{X}_{k,+1}} p(Y^{(b)}|Q_X)P(Q_X)}{\sum_{Q_X \in \mathcal{X}_{k,-1}} p(Y^{(b)}|Q_X)P(Q_X)}, \tag{5.2}$$

where $\mathcal{X}_{k,\pm 1}$ is the set of all matrices Q_X in the constellation having $x_k = \pm 1$. That is,

$$\mathcal{X}_{k,\pm 1} = \{Q_X \in \mathcal{C} | x_k = \left[\mathcal{M}^{-1}(Q_X)\right]_k = \pm 1\}, \tag{5.3}$$

where $[\cdot]_k$ denotes the k^{th} element of the vector. The likelihood of the received signal given the transmitted unitary matrix is [12, 40]

$$p(Y|Q_X) = \frac{\exp\left(-\frac{\rho T}{M}\mathrm{Tr}(Y^\dagger(I_T - \frac{1}{1+M/\rho T}Q_X Q_X^\dagger)Y)\right)}{(\pi M/\rho T)^{TN}(1+\rho T/M)^{MN}}. \tag{5.4}$$

If the interleaver in Figure 5.1 is well designed, the n bits that label Q_X are approximately independent and hence we can approximate $P(Q_X)$ by

$$P(Q_X) \approx \prod_{k=1}^{n} P(x_k^{(b)} = [\mathcal{M}^{-1}(Q_X)]_k). \tag{5.5}$$

In the first iteration, we assume that the bits are equi-probable and hence $P(Q_X) = 1/2^n$. In subsequent iterations, the components on the right hand side of (5.5) are computed from the interleaved decoder output from the previous iteration; namely $L_{A_1}(x_k) = \ln \frac{P(x_k=+1)}{P(x_k=-1)}$. The final step in the demapping process is to extract the extrinsic information L_{E_1} by subtracting the *a priori* information L_{A_1} from the demapper output L_{D_1}; see Figure 5.1. This extrinsic information is then deinterleaved to act as the *a priori* input L_{A_2} for the soft-input/soft-output outer decoder. As shown in Section 4.4, the turbo decoder uses the BCJR algorithm [39] to calculate the *a posteriori* L-values for the coded bits for each constituent code. The *a posteriori* L-values of the constituent codes are then concatenated as described in Section 4.4.2, generating L_{D_2}. The *a priori* information L_{A_2} is subtracted from L_{D_2} resulting in the extrinsic soft information L_{E_2}. This extrinsic information L_{E_2} is then interleaved to act as the *a priori* information L_{A_1} for the next iteration of the demapper; completing an iteration.

5.3.1 List-Based Demapper

For most implementations of the BICM scheme shown in Figure 5.1, the computational bottleneck will be the computation of the *a posteriori* L-value (5.2) in the demapper. In the case of coherent MIMO communication systems, it was shown in [16, 51] that the impact of this bottleneck can be mitigated without a significant loss in performance by approximating $L_{D_1}(x_k^{(b)}|Y^{(b)})$ in (5.2) by computing the summation on the right hand side of (5.2) over a list of candidate constellation points rather than over the whole constellation. In order to apply the principles of that approach in the non-coherent case, the Grassmannian geometry of the constellation and the information in the received signal $Y^{(b)}$ has to be properly exploited to identify a list $\mathcal{L}_b$ of candidate constellation points. The *a posteriori* L-value in (5.2) can be

approximated by

$$L_{D_1}(x_k^{(b)}|Y^{(b)}) \approx \ln \frac{\sum_{Q_X \in \mathcal{X}_{k,+1}^R} p(Y^{(b)}|Q_X)P(Q_X)}{\sum_{Q_X \in \mathcal{X}_{k,-1}^R} p(Y^{(b)}|Q_X)P(Q_X)}, \tag{5.6}$$

where $\mathcal{X}_{k,\pm 1}^R$ is the set of matrices in the list for which $x_k = \pm 1$; i.e.,

$$\mathcal{X}_{k,\pm 1}^R = \{Q_X \in \mathcal{L}_b | x_k^{(b)} = \left[\mathcal{M}^{-1}(Q_X)\right]_k = \pm 1\}. \tag{5.7}$$

The proposed demapping list $\mathcal{L}_b$ will be based on the list generated by the reduced-search non-coherent detector for the uncoded Grassmannian constellations developed in [27, 36, 37]. As described in Section 2.4.1, in this list generation technique, the channel output is exploited to generate the list of candidate points. A desirable feature of this list demapper it that the length of this list is adapted to the channel realization, allowing the receiver to allocate its computational resources to channel realizations in which the detection process is hard. In the next section, we will present a simple technique by which the decoder's soft information can be used to augment the demapper's list, and obtain significant performance gain.

5.3.2 List Augmentation for the List-based Demapper

In the direct application of the reduced-search non-coherent MIMO detector in Section 2.4.1 to list-based demapping, membership of the list is determined entirely by the channel output. A weakness in applying this strategy in a BICM-IDD scheme with long (outer) codewords is that a constellation point whose binary index is deemed by the decoder to have a large likelihood might not be a member of the demapper's list. If that were to occur, the system would not benefit from the flow of soft information between the demapper and the decoder.

To mitigate this effect, we propose to incorporate the decoder's soft information in the list construction algorithm. A computationally efficient way to do this is to simply

augment the demapper's list at each iteration with the constellation point whose binary index is deemed by the decoder to have the largest likelihood. To describe the augmentation process, we let $L_{A_1}^{[i]}$ denote the vector of *a priori* information used by the demapper in the i^{th} demapping-decoding iteration; i.e., $L_{A_1}^{[i]}$ is the interleaved version of $L_{E_2}^{[i-1]}$, the extrinsic information from the decoder at the end of the $(i-1)^{th}$ iteration; see Figure 5.1. Before performing the list-based demapping at the i^{th} iteration, the demapper makes an (auxiliary) hard decision on the b^{th} length-n block of $L_{A_1}^{[i]}$ and checks whether the constellation point that corresponds to that hard decision is on the demapper's current list of candidate points. If it is not on the list, this constellation point is added to the list.

For a received matrix $Y^{(b)}$, the augmented list of candidate constellations for the i^{th} iteration can be written as

$$\mathcal{L}_b^{[i]} = \mathcal{L}_b^{[i-1]} \bigcup \left\{ \mathcal{M}\left(\operatorname{sgn}\left(\left(L_{A_1}^{[i]}(x_k^{(b)}) \right)_{k=1}^{n} \right) \right) \right\}, \tag{5.8}$$

and hence, at the i^{th} iteration the L-value of the k^{th} coded bit in the b^{th} block is approximated using

$$L_{D_1}^{[i]}(x_k^{(b)}|Y^{(b)}) \approx \ln \frac{\sum_{Q_X \in \mathcal{X}_{k,+1}^{R^{[i]}}} p(Y^{(b)}|Q_X)P(Q_X)}{\sum_{Q_X \in \mathcal{X}_{k,-1}^{R^{[i]}}} p(Y^{(b)}|Q_X)P(Q_X)}, \tag{5.9}$$

where $\mathcal{X}_{k,\pm 1}^{R^{[i]}}$ is the set of matrices Q_X in the augmented candidate list for $x_k = \pm 1$,

$$\mathcal{X}_{k,\pm 1}^{R^{[i]}} = \{Q_X \in \mathcal{L}_b^{[i]} | x_k^{(b)} = \left[\mathcal{M}^{-1}(Q_X)\right]_k = \pm 1\}. \tag{5.10}$$

As illustrated in Figure 5.2, allowing the decoder to augment the demapper's list in this way offers the potential for significant performance improvements. Furthermore, since we add at most one constellation point to the demapper's list at each iteration, the increase in demapper complexity is negligible when compared to the computational cost that would be incurred if all the constellation points were considered, as in (5.2). There are other ways in which the decoder could augment the

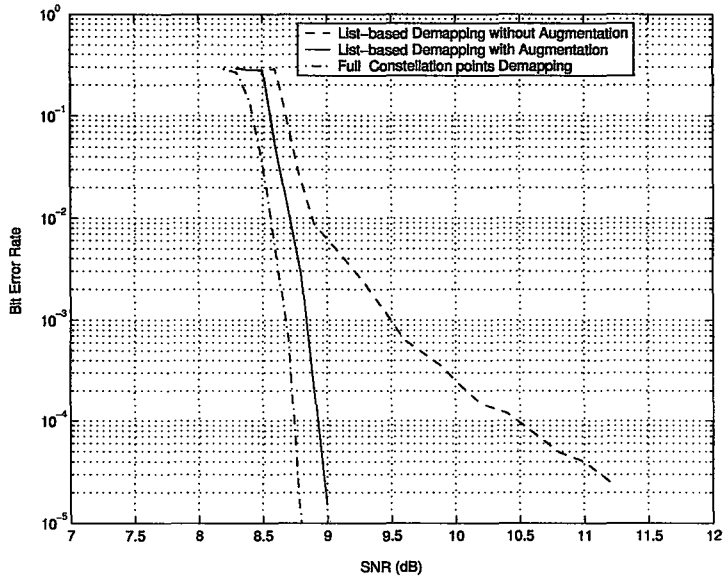


Figure 5.2: Bit error rate performance of the proposed BICM-IDD scheme using full demapping (as in (5.2), dash-dot), and using list-based demapping with (solid) and without (dashed) list augmentation. The transmission scheme has the rate 1/2 turbo outer code and the constellation of 256 unitary matrices described in Section 5.5.

demapper's list, such as by examining the unreliable entries in $L_{A_1}^{[i]}$ and adding the constellation pairs that correspond to those entries. However, in the example shown in Figure 5.2 the proposed low-complexity augmentation scheme appears to extract much of the potential gain.

5.4 BICM-IDD Parameter Selection

There are several factors that can affect the performance of a BICM-IDD scheme. As described in Section 3.2, the proposed BICM-IDD scheme consists of an outer and inner phase. The outer phase consists of the outer encoder and its corresponding soft-input soft-output decoder. In the presented simulations, a standard turbo encoder and decoder pair was used. The decoder iteratively employs the BCJR algorithm to update the *a posteriori* probability for each bit. The considered turbo encoder

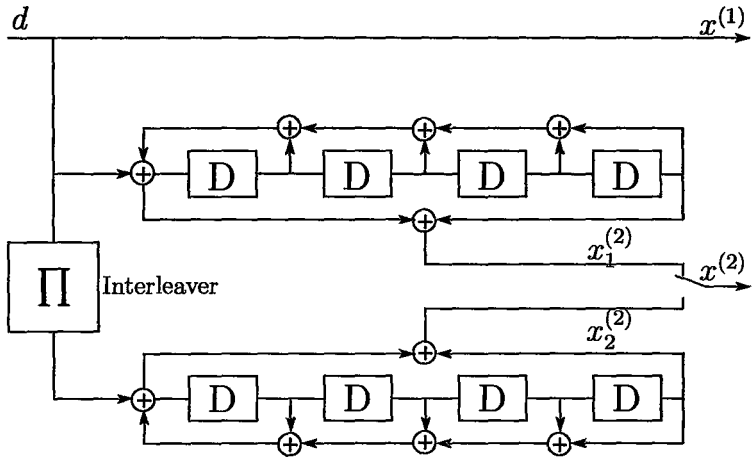


Figure 5.3: A rate 1/2 punctured recursive systematic turbo code.

consisted of two memory-4 recursive systematic convolutional codes concatenated in parallel. This choice was motivated by the choice made in [52]. A block diagram of the encoder is shown in Figure 5.3. As discussed in Section 4.4, the turbo decoder iteratively exchanges soft information between its constituent BCJR decoders for several iterations, which presents a trade-off between the complexity and the performance of the decoding process. In [16], it was observed that increasing the number of turbo iterations beyond 8 results in an increase in the decoding complexity without significantly improving the performance. Motivated by that observation, in the proposed scheme the number of turbo iterations was fixed *a priori* to 8.

The inner phase of the BICM-IDD scheme consists of the unitary constellation mapper and the proposed list-based demapper. The mapper maps consecutive blocks of bits onto unitary constellations of size 256 or 1024. The constellations were generated using the greedy algorithm in [27] that was described in Section 2.3. The chosen constellation points are then transmitted from 2 transmit antennas to 2 receive antennas over a frequency-flat richly-scattered block-fading channel with a signalling interval $T = 4$. At the receiver, the list-based demapper depends on several factors, namely, the number of reference points and the width of the strap associated with

each reference point. Such parameters result in a trade-off between the performance and complexity of the list-based demapper and will be addressed in Section 5.4.2. In addition, the use of the list-based demapper rather than full constellation demapping may result in bit positions in which all the bit sequences in the candidate list have the same binary value. This indicates that the L-values associated with those bits will be ∞ or $-\infty$ since the denominator or the numerator of the L-value will be zero, respectively; c.f., (5.6). Therefore, a positive and negative clipping value $\pm L_{\text{clip}}$ will be assigned [53]. Section 5.4.3 will address the choice of the clipping value. In the BICM-IDD receiver, the list-based demapper and the turbo decoder iteratively exchange soft information to approximate the prohibitively complex bit-wise MAP decoder. The choice of the number of demapper to decoder iterations will be considered in the next section.

5.4.1 Choice of Number of Iterations Between Demapper and Decoder

The number of iterations of soft information exchange between the demapper and the decoder presents a trade-off between the performance and computational cost of the detection process in the BICM-IDD scheme. In Figure 5.4, the performance of non-coherent MIMO BICM-IDD schemes that use two, four, and eight demapper to decoder iterations are presented. As shown in Figure 5.4, the performance of the system improves significantly as the number of iterations is increased from two to four, while the performance improvement of having eight iterations rather than four is rather small when balanced against the complexity associated with doubling the number of iterations. Thus, in the simulations provided in this thesis, the number of demapper to decoder iterations is set to four.

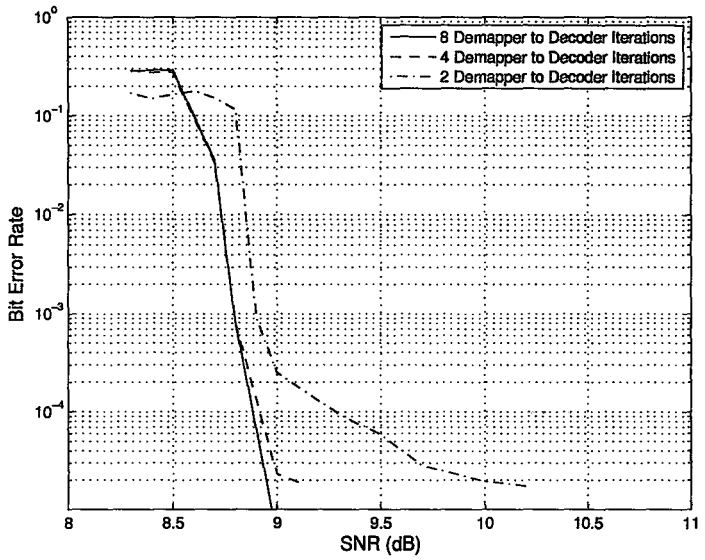


Figure 5.4: Bit error rate performance of the proposed BICM-IDD scheme using 2 demapper to decoder iterations (dash-dot), 4 demapper to decoder iterations (dashed) and 8 demapper to decoder iterations (solid).

5.4.2 Choice of List Demapper Parameters

In the list-based demapper, the number of reference points is chosen by considering the tradeoff between the performance and the complexity of the demapper. As described in Section 2.4.1, increasing the number of reference points result is shorter lists since the list candidate points are only the constellation points that lie in the intersection of the straps. However, the reduction in the search space increases the probability that the transmitted constellation point is not in the list. Therefore, it was proposed in [36, 37] to increase the width of the strap associated with each reference point as the number of reference points increases. It was also shown in [36, 37], that if the width of the strap is appropriately adjusted, increasing the number of reference points improves the performance of the list demapper while reducing the number of likelihood computations. However, the drawback of increasing the number of reference points is the memory requirement of the receiver in order to store all the distances between

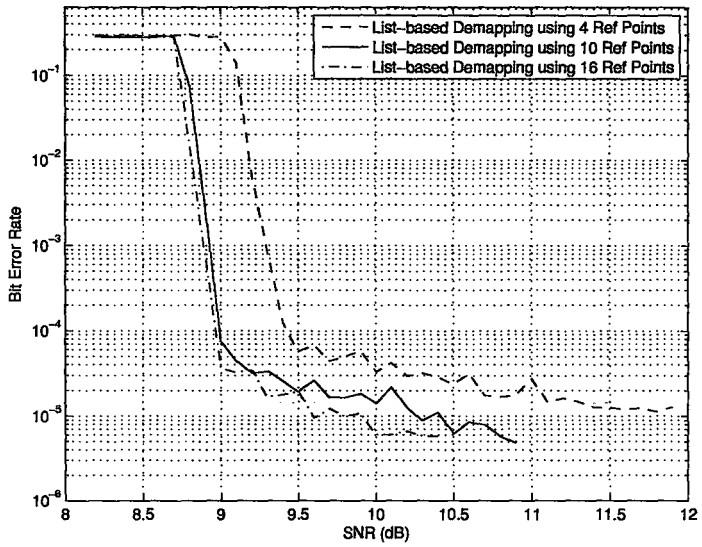


Figure 5.5: Bit error rate performance of the proposed BICM-IDD scheme in which the list demapper using 4 reference points (dashed), 10 reference points (solid) and 16 reference points (dash-dot).

all the constellation points and each reference point and the overhead complexity associated with computing all the distances.

In Figure 5.5, the bit error performance of the non-coherent BICM-IDD scheme is presented for a system using 4, 10, and 16 demapper reference points. As shown, increasing the number of reference points improves the performance of the BICM-IDD system. On the other hand in Figure 5.6, the average number of likelihood computations are shown. This figure shows that the number of candidate points in the list-based demapper is not significantly reduced as the number of reference points increases. Since the performance and complexity gap between using 10 and 16 reference points is rather small, both 10 and 16 reference points would be reasonable choices for the simulations. However, in order to reduce the memory required by the receiver and the overhead complexity, we chose to use 10 reference points in the presented simulations.

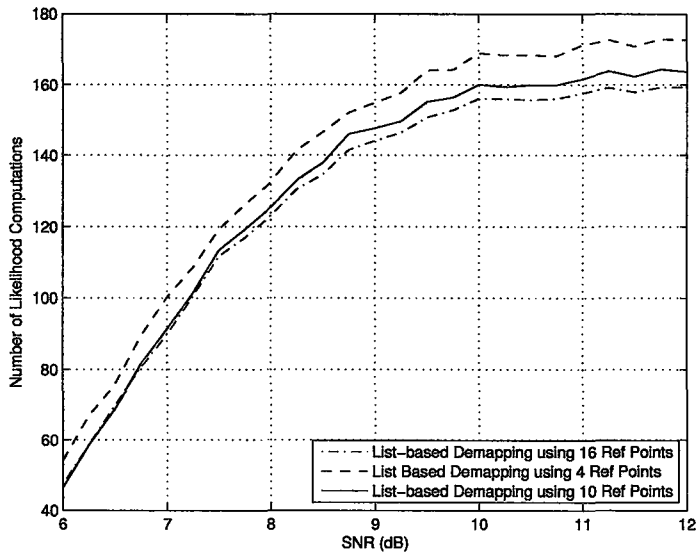


Figure 5.6: The average number of likelihood computations for the list demapper using 4 reference points (dashed), 10 reference points (solid) and 16 reference points (dash-dot).

In order to significantly reduce the number of likelihood computations of the list demapper, one might choose to reduce the strap width associated with each reference point. Alternatively, one might choose to increase the strap width to improve the performance of the non-coherent BICM-IDD system. In Figure 5.7, the associated complexity of the likelihood computations for the list demapper is illustrated, where the strap width is increased and decreased by 20 percent overthat recommended in [36, 37]. Figure 5.8 presents the performance of a BICM-IDD scheme using the increased and decreased strap width. As shown, reducing the strap width results in a considerably shorter list but significantly degrades the performance of the system. On the contrary, increasing the strap width results in a significantly longer list while only providing a slight improvement in performance.

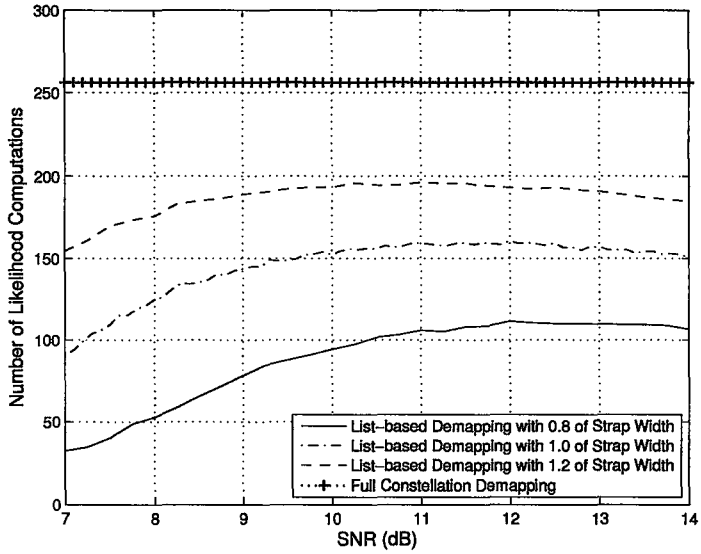


Figure 5.7: The average number of likelihood computations for the list demapper with 0.8 of strap width (solid), 1.0 of strap width (dash-dot), 1.2 of strap width (dashed) and full constellation demapping (+).

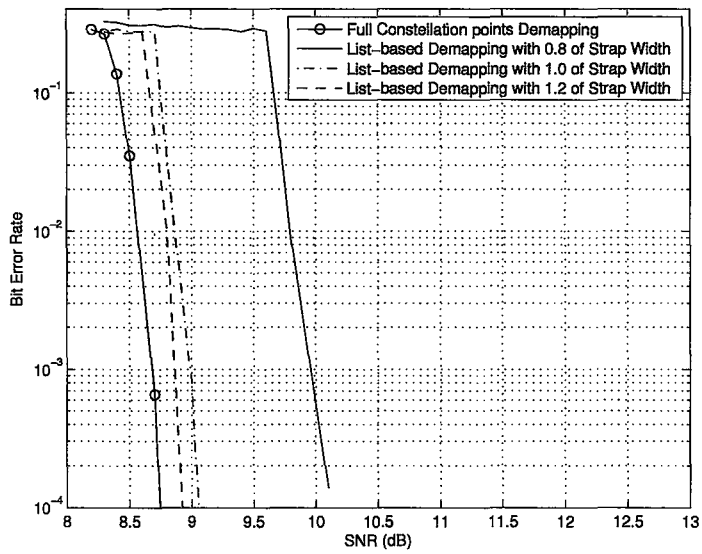


Figure 5.8: Bit error rate performance of the proposed BICM-IDD scheme in which the list demapper with 0.8 of strap width (solid), 1.0 of strap width (dash-dot), 1.2 of strap width (dashed) and full constellation demapping (o).

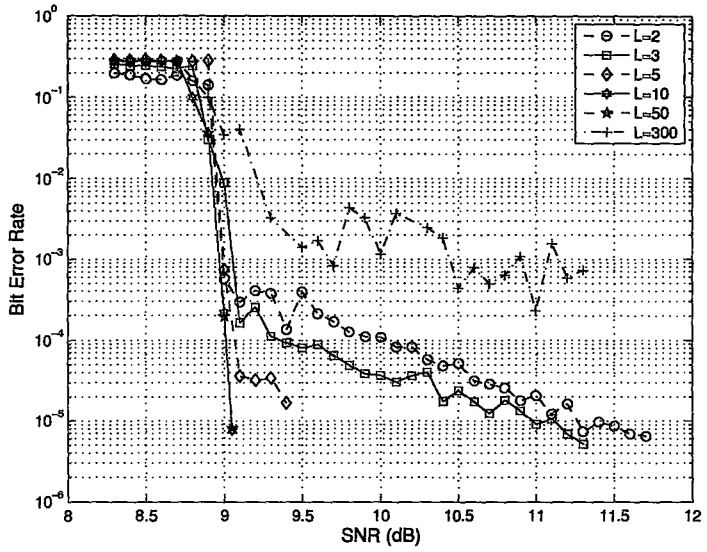


Figure 5.9: Bit error rate performance of the proposed BICM-IDD scheme using different clipping values.

5.4.3 Choice of Clipping Value

It was shown in [53] that the clipping value L_{clip} can affect the error performance of the iterative receiver. A clipping value that is much higher than the optimum value forces the decoder to assume that the clipped values from the demapper have the correct sign leading the decoder to ignore soft information associated with other bits [53]. On the other hand, a much lower clipping value results in a significant loss in the demapper's soft information, which causes the demapper to have no sufficient influence on the decoder [53]. In Figure 5.9, the performance of the BICM scheme for different clipping values L_{clip}. As shown a fixed clipping value $L_{clip} = \pm 10$ yielded consistently good results.

5.5 Performance Simulations of Unitary Signalling Versus Training-based Technique

In this section, the performance of the proposed BICM-IDD unitary signalling scheme will be compared to that of a corresponding BICM-IDD training scheme.

As discussed in Section 2.5, an alternative to the unitary signalling approach to non-coherent MIMO communication is a training-based approach [3, 32, 33]. The training-based approach splits the signalling interval T into two subintervals. The first subinterval is a training phase in which pilot symbols known to the receiver are transmitted and an estimate $\hat{H}$ of the channel matrix is determined by the receiver. In the second subinterval, the receiver uses the estimated channel matrix $\hat{H}$ to detect the transmitted data coherently.

We consider a system with data rates of 1, 5/4 and 5/3 bits per channel use (bpcu). In the unitary signalling case, the data rate is given by $\frac{R\log_2|C|}{T}$ where R is the rate of the outer encoder, $|C|$ is the size of the unitary constellation, and T is the signalling interval in channel uses. The data rate of 1 bpcu scheme consists of a constellation of 256 unitary matrices and an outer code of rate 1/2, and the rate 5/4 and 5/3 bpcu schemes consist of a constellation of 1024 unitary matrices and outer codes of rates 1/2 and 2/3 respectively. In all these cases, the data signalling interval $T = 4$. A block diagram of the rate 1/2 encoder is shown in Figure 5.3 and the rate 2/3 encoder is that shown in [54]. The constellations of unitary matrices were designed using the greedy algorithm in [27].

In the training case described in Section 2.5, two channel uses were used for the training, i.e., $T_\tau = 2$, and the Alamouti scheme [43] used two channel uses for the data communication phase; i.e., $T_d = 2$. The data rate is given by $\frac{RZ\log_2|C_s|}{T_\tau+T_d}$, where Z is the number of scalar symbols per space-time codeword and $|C_s|$ is the size of the standard scalar constellation. For the Alamouti scheme, $Z = 2$. The data rate

of 1 bpcu was achieved using the Gray-labelled 16-QAM constellation along with the rate 1/2 outer code, and the rate 5/4 and 5/3 bpcu schemes consisted of the Gray-labelled 32-cross-QAM constellation along with the outer codes of rates 1/2 and 2/3, respectively. In the training scheme the formula for the data rate is slightly different from that of the unitary scheme due to the fact that in the unitary scheme a unitary matrix of size 4×2 is transmitted in each 4 channel uses while in the training scheme a pilot matrix of size 2×2 and a data matrix of size 2×2 are transmitted in each 4 channel uses.

As discussed in Section 5.4 and as is apparent from Figure 5.3, the outer codes are systematic parallel concatenated turbo codes [52] with recursive convolutional codes of memory length 4 as the constituent codes. To employ these codes in a blockwise fashion, the data to be transmitted was partitioned into blocks of length 8000. Based on the simulation results in Section 5.4, at the receiver, four demapping-decoding iterations were performed for each block, with eight "turbo" iterations being performed within the outer decoder for each demapping-decoding iteration. The demapper used 10 reference points to compute the candidate list, and in both the demapper and decoder the L-values were clipped at ± 10.

As shown in Figure 5.10, in this example the training-based BICM-IDD scheme performs better than the proposed unitary-signalling-based BICM-IDD scheme at low data rates. At a data rate of 1 bpcu, training performs better than the proposed scheme by 1.8 dB at a bit error rate (BER) of 10^{-5}. If the data rate is increased to 5/4 bpcu, the performance gap is decreased to 0.2 dB at a BER of 10^{-5}. However, at higher data rates the unitary signalling scheme outperforms the training scheme. At a data rate of 5/3 bpcu the SNR gain of the unitary scheme at a BER of 10^{-5} is close to 1 dB. The proposed scheme performs better than the training-based scheme at higher data rates because the channel symbols of the proposed scheme mimic the distribution that achieves capacity at high SNRs. In particular, the training scheme only achieves

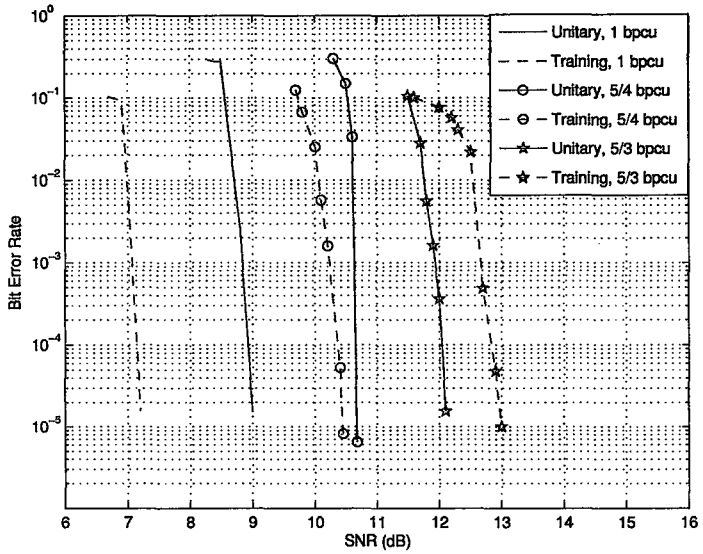


Figure 5.10: Bit error rate performance of the proposed BICM-ID scheme (solid) and a training-based BICM-ID scheme (dashed) for data rates of 1 (no symbol), 5/4 ($\circ$) and 5/3 ($\star$) bits per channel use.

the SNR-dependent term in the high SNR non-coherent ergodic capacity in (2.2), whereas the unitary signalling scheme achieves (2.2) including the SNR-independent constant $c_{M,N}$. This constant can be substantial, especially in the case where the number of receive antennas exceeds the number of transmit antennas [36]. The results in Figure 5.10 show that even when the number of antennas at each end of the link is equal (to 2), employing a signalling scheme that exploits the SNR-independent term can lead to significant gains at high data rates.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis, we proposed a BICM-IDD scheme for non-coherent MIMO systems that provides reliable communication at high data rates. This scheme was developed using the observation that the channel symbols that achieve the non-coherent ergodic capacity at high SNRs can be represented as constellation points on a compact Grassmann manifold. The transmission scheme was based on an existing constellation design that mimics the capacity achieving distribution. In order to incorporate the BICM-IDD scheme a mapping technique that exploits some of the properties of that constellation was developed. In addition, in order to alleviate the computational burden of examining all the constellation points, a list-based demapper was developed.

The complexity of directly generating a constellation that mimics the capacity achieving distribution of the non-coherent MIMO channel grows rapidly with the size of the constellation. Thus, a greedy algorithm with a smoothed objective was used to generate a well-spaced constellation. In the non-coherent BICM-IDD scheme, a labelling strategy is required in order to map consecutive blocks of bits onto the

Grassmannian constellation. Hence, a constellation mapper that exploits the Grassmannian structure of the signal was developed based on incomplete set partitioning.

In common with BICM schemes for the coherent MIMO channel, the computational burden of the receiver usually lies in the demapper, in which complete enumeration over all the constellation points is performed to generate the *a posteriori* soft information. An efficient list-based demapper that approximates the *a posteriori* soft information by performing the enumeration over a subset of constellation points was proposed. The list-based demapper utilizes certain properties of the Grassmannian geometry as well as the nature of the received signal to generate a list of candidate points.

A weakness in this list-based demapper was that the list of candidate constellation points is entirely determined by the channel output without any input from the decoder of the BICM-IDD scheme. In other words, a constellation point that is deemed by the decoder to have a large likelihood might not be a member of the demapper's candidate list. Thus, it was proposed to incorporate the decoder's information in the demapper by allowing the decoder to augment the demapper's candidate list in each detection iteration. We demonstrated that the proposed augmentation scheme resulted in a significant improvement in the performance with negligible added complexity.

Finally, the performance of the proposed BICM-IDD unitary signalling scheme was compared to a corresponding training based scheme. We demonstrated that at high data rates the proposed BICM scheme outperforms the training-based scheme.

6.2 Future Work

There are several credible approaches for improving the already promising performance of the proposed BICM-IDD non-coherent scheme. In particular, the proposed

constellation mapping was based on a simple incomplete set partitioning that partitions the constellations to two subsets. It is expected that better insight into the geometry of (imperfect) Grassmannian lattices will lead to mappings with improved performance.

In the detection of the proposed scheme, the number of iterations between the demapper and the decoder in which extrinsic soft information is exchanged was set *a priori*. Adaptively controlling the number of iterations between the demapper and the decoder offers the potential to improve the trade-off between performance and complexity of the receiver.

The decoder of the constituent codes that was used to construct the outer code was a standard BCJR algorithm. The complexity of this algorithm can be significantly reduced by discarding certain branches in the trellis based on the input soft information from the demapper. Since the proposed scheme is using a list-based demapper, it might occur that for a certain bit position all the bit sequences that corresponds to the constellation points in the candidate list have the same binary value. This fact can be exploited in the decoder by only considering the trellis branches associated with this binary value and discarding the others. Further reduction in complexity can be achieved by examining the soft information from the demapper and discarding all the branches associated with small probabilities.

In the proposed BICM scheme, the decoder augments the demapper's candidate list by at most one constellation point in each iteration. Combinations of more sophisticated list augmentation strategies and possible list reduction strategies might (uniformly) improve the trade-off between performance and computational cost provided by the current demapper.

Bibliography

[1] S. Barbarossa, *Multiantenna Wireless Communication Systems.* Boston: Artech House, 2005.

[2] J. G. Proakis, *Digital Communications.* New York: McGraw-Hill, 3rd ed., 1995.

[3] B. Hassibi and B. Hochwald, "How much training is needed in multiple-antenna wireless links?," *IEEE Trans. Inform. Theory*, vol. 49, pp. 951–963, Apr. 2003.

[4] J. R. Barry, E. A. Lee, and D. G. Messerschmitt, *Digital Communication.* Boston: Kluwer Academic, 3rd ed., 2004.

[5] R. R. Muller and W. H. Gerstacker, "On the capacity loss due to seperation of detection and decoding," *IEEE Trans. Inform. Theory*, vol. 50, pp. 1769–1778, Aug. 2004.

[6] S. Shamai and R. Laroia, "The intersymbol interference channel: lower bounds on capacity and channel precoding loss," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1388–1404, Sept. 1996.

[7] L. Zheng and D. N. C. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple antenna channels," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1073–1096, May 2003.

[8] K. Azarian and H. El Gamal, "The throughput-reliability tradeoff in mimo channels," Sept. 2005. Accepted for publication subject to revesions in the *IEEE Trans. Inform. Theory.* Available online: `http://arxiv.org/abs/cs.IT/0509021`.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.

[10] I. E. Telatar, "Capacity of multiantenna Gaussian channels," *Eur. Trans. Telecom.*, vol. 10, pp. 585–595, Nov. 1999.

[11] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE J. Select. Areas Commun.*, vol. 21, pp. 684–701, June 2003.

[12] L. Zheng and D. N. C. Tse, "Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Trans. Inform. Theory*, vol. 48, pp. 359–383, Feb. 2002.

[13] E. Zehavi, "8-PSK trellis codes for a Rayleigh channel," *IEEE Trans. Commun.*, vol. 40, pp. 873–884, May 1992.

[14] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Trans. Inform. Theory*, vol. 44, pp. 927–945, May 1998.

[15] X. Li and J. A. Ritcey, "Bit-interleaved coded modulation with iterative decoding using soft feedback," *IEE Elect. Lett.*, vol. 34, pp. 942–943, May 1998.

[16] B. M. Hochwald and S. ten Brink, "Achieving near-capacity on a multiple-antenna channel," *IEEE Trans. Commun.*, vol. 51, pp. 389–399, Mar. 2003.

[17] H. Imai and S. Hirakawa, "A new multilevel coding method using error correcting codes," *IEEE Trans. Inform. Theory*, vol. 23, pp. 371–377, May 1977.

[18] U. Wachsmann, R. F. C. Fischer, and J. B. Huber, "Multilevel codes: Theoretical concepts and practical design rules," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1361–1391, July 1999.

[19] L. H. J. Lampe, R. Schober, and R. F. C. Fischer, "Multilevel coding for multiple antenna transmission," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 203–208, Jan. 2004.

[20] I. Bahceci and T. M. Duman, "Combined turbo coding and unitary space-time modulation," *IEEE Trans. Commun.*, vol. 50, pp. 1244–1249, Aug. 2002.

[21] A. Sezgin, E. A. Jorswieck, and H. Boche, "On EXIT-chart analysis of coherent and non-coherent space-time codes," in *Proc. ITG Workshop Smart Antennas*, (Munich), pp. 49–56, Mar. 2004.

[22] T. Li, W. H. Mow, and K. B. Letaief, "Bit-interleaved coded modulation for noncoherent iterative decoding for multiple antenna systems," in *Proc. Int. Conf. Communications, Circuits and Systems*, vol. 1, (Chengdu, China), pp. 135–139, June 2004.

[23] Y. Li and X. G. Xia, "Constellation mapping for space-time matrix modulation with iterative demodulation/decoding," *IEEE Trans. Commun.*, vol. 53, pp. 764–768, May 2005.

[24] M. L. McCloud, M. Brehler, and M. K. Varanasi, "Signal design and convolutional coding for noncoherent space-time communication on the block-Rayleigh-fading channel," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1186–1194, May 2002.

[25] D. Agrawal, T. Richardson, and R. Urbanke, "Multiple-antenna signal constellations for fading channels," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2618–2626, Sept. 2001.

[26] J. H. Conway, R. H. Hardin, and N. J. A. Sloane, "Packing lines, planes, etc.: Packings in Grassmannian spaces," *Exper. Math.*, vol. 5, no. 2, pp. 139–159, 1996.

[27] R. H. Gohary and T. N. Davidson, "Non-coherent MIMO communication: Grassmannian constellations and efficient detection," in *Proc. IEEE Int. Symp. Inform. Theory*, (Chicago, USA), June 2004.

[28] B. M. Hochwald, T. L. Marzetta, T. J. Richardson, W. Sweldens, and R. L. Urbanke, "Systematic design of unitary space-time constellations," *IEEE Trans. Inform. Theory*, vol. 46, pp. 1962–1973, Sept. 2000.

[29] B. M. Hochwald and T. L. Marzetta, "Unitary space-time modulation multiple-antenna communications in Rayleigh flat fading," *IEEE Trans. Inform. Theory*, vol. 46, pp. 543–564, Mar. 2000.

[30] B. Hassibi and B. M. Hochwald, "Cayley differential unitary space-time codes," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1485–1503, June 2002.

[31] A. L. Moustakas, S. H. Simon, and T. L. Marzetta, "Capacity of differential versus nondifferential unitary spacetime modulation for MIMO channels," *IEEE Trans. Inform. Theory*, vol. 52, pp. 3622–3634, Aug. 2006.

[32] P. Dayal, M. Brehler, and M. K. Varanasi, "Leveraging coherent space-time codes for noncoherent communication via training," *IEEE Trans. Inform. Theory*, vol. 50, pp. 2058–2080, Sept. 2004.

[33] H. El Gamal, D. Aktas, and M. O. Damen, "Noncoherent space-time coding: An algebraic perspective," *IEEE Trans. Inform. Theory*, vol. 51, pp. 2380–2390, July 2005.

[34] Y. Jia, C. Andrieu, R. J. Piechocki, and M. Sandell, "Joint channel tracking and symbol detection in MIMO systems via multiple model methods," in *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications*, (New York), pp. 12–16, June 2005.

[35] G. Taricco and E. Biglieri, "Space-time decoding with imperfect channel estimation," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 1874–1888, July 2005.

[36] R. H. Gohary, *Efficient Space-Time Signalling: Coherent and Non-Coherent Scenarios.* PhD thesis, McMaster University, Apr. 2006.

[37] R. H. Gohary and T. N. Davidson, "On efficient non-coherent detection of Grassmannian constellations," in *Proc. IEEE Int. Symp. Inform. Theory*, (Adelaide, Australia), pp. 1676–1680, Sept. 2005.

[38] C. Berrou and A. Glavieux, "Near optimum error correcting coding and decoding: Turbo-codes," *IEEE Trans. Commun.*, vol. 44, pp. 1261–1271, Oct. 1996.

[39] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inform. Theory*, vol. 20, pp. 284–287, Mar. 1974.

[40] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading," *IEEE Trans. Inform. Theory*, vol. 45, pp. 139–157, Jan. 1999.

[41] M. J. Borran, A. Sabharwal, and B. Aazhang, "On design criteria and construction of noncoherent space-time constellations," *IEEE Trans. Inform. Theory*, vol. 49, pp. 2332–2351, Oct. 2003.

[42] B. Hassibi and B. M. Hochwald, "High-rate codes that are linear in space and time," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1804–1824, July 2002.

[43] S. M. Alamouti, "A simple transmitter diversity scheme for wireless communications," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 1451–1458, Oct. 1998.

[44] J. Hagenauer, E. Offer, and L. Papke, "Iterative decoding of binary block and convolutional codes," *IEEE Trans. Inform. Theory*, vol. 42, pp. 429–445, Mar. 1996.

[45] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inform. Theory*, vol. 28, pp. 55–67, Jan. 1982.

[46] J. Sun and O. Y. Takeshita, "Interleavers for turbo codes using permutation polynomials over integer rings," *IEEE Trans. Inform. Theory*, vol. 51, pp. 101–119, Jan. 2005.

[47] L. C. Perez, J. Seghers, and D. J. Costello Jr., "A distance spectrum interpretation of turbo codes," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1968–1709, Nov. 1996.

[48] O. Y. Takeshita and D. J. Costello Jr., "New deterministic interleaver designs for turbo codes," *IEEE Trans. Inform. Theory*, vol. 46, pp. 1988–2006, Sept. 2000.

[49] N. Gresset, J. J. Boutros, and L. Brunel, "Multidimensional mappings for iteratively decoded BICM on multiple-antenna channels," *IEEE Trans. Inform. Theory*, vol. 51, pp. 3337–3346, Sept. 2005.

[50] Z. Hong and B. L. Hughes, "Bit-interleaved space-time coded modulation with iterative decoding," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 1912–1917, Nov. 2004.

[51] H. Vikalo, B. Hassibi, and T. Kailath, "Iterative decoding for MIMO channels via modified sphere decoding," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 2299–2311, Nov. 2004.

[52] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes," in *Proc. Int. Conf. Commun.*, (Geneva, Switzerland), pp. 1064–1070, May 1993.

[53] Y. L. C. de Jong and T. J. Willink, "Iterative tree search detection for MIMO wireless systems," *IEEE Trans. Commun.*, vol. 53, pp. 930–935, June 2005.

[54] S. Benedetto, R. Garello, and G. Montorsi, "A search for good convolutional codes to be used in the construction of turbo codes," *IEEE Trans. Commun.*, vol. 46, pp. 1101–1105, Sept. 1998.