Low-Complexity Detection and Precoding in High Spectral Efficiency Large-MIMO Systems

A Thesis
Submitted for the Degree of
Doctor of Philosophy
in the Faculty of Engineering

Submitted by
Mohammed Saif Khan

Electrical Communication Engineering
Indian Institute of Science, Bangalore
Bangalore – 560 012 (INDIA)

March 2010
Acknowledgments

My Ph.D has been made possible not just due to me, but support and motivation from everybody around me. I would firstly thank my parents, for instilling good values in me. Being from a humble background, my parents taught me to always strive hard and never give up. I would also thank my wife Saba and my daughter Samaira for supporting me throughout my Ph.D. I would specially thank my wife for taking over all my personal responsibilities so that I could focus on my Ph.D. It was my wife, who believed in my capabilities more than me, and who convinced me to continue my education. I would also like to say sorry to my little 3 year old daughter, with whom I could not spend as much time.

Prof. A. Chockalingam has been my mentor and adviser, without whom this Ph.D was not possible. Apart from sharing his vast knowledge and experience in the field of communications, he has also taught me the art of research. Specifically, I do remember that initially I used to get stuck with problems and would waste time without getting a solution. He, then advised me to look at the smaller version of the problem and then attack the original problem. This mantra has worked for me at many occasions during my Ph.D. There were many more mantras which he has taught me, and without which I could not have finished my Ph.D. I also admire him for the freedom he gives to think and timely support, which made my Ph.D a very pleasurable experience which I would cherish all through my entire life.

I would also like to express my gratitude and thanks to Prof. B. Sundar Rajan, for giving his valuable time and technical support for most of my Ph.D. I still remember, the extended late evening technical discussions on large-MIMO that all three of us used to have in the coffee house. I cannot forget thanking Prof. Emanuele Viterbo at the University of Calabria, Italy, with whom I had spent the last six months of my
Ph.D, and without whom the two Chapters on Large-MIMO precoding would not have materialized. I specially thank Prof. E. Viterbo for always being available for timely discussions despite his busy schedule, and also for taking care of my logistics which made my stay very pleasant.

I would also thank my parents in-law, my brother and sister, for having supported me and being besides me in my lows and highs. I would also make a special mention of all my lab mates and the friendly technical discussions which had a positive bearing on my Ph.D. I would thank the administrative staff of the ECE Dept. and IISc, without whose logistical support my Ph.D would not have materialized. Finally, I cannot forget thanking the almighty, for having made me stronger and perseverant from inside and having bestowed blessings on me.
Abstract

The research reported in this thesis is concerned with multiple-input multiple-output (MIMO) systems that employ large number of transmit/receive antennas. MIMO systems with tens of antennas in communication terminals, referred to as large-MIMO systems, are considered. The motivation to consider such large-MIMO systems is the potential to practically realize the theoretically predicted benefits of MIMO, in terms of both high spectral efficiencies as well as increased diversity orders, through the exploitation of large spatial dimensions. High complexity of detection and precoding in such large-MIMO systems has been a major issue. This thesis focuses on the design of large-MIMO detection and precoding algorithms that can achieve near-optimal performance at practically affordable low complexities. The work reported in the thesis is comprised of the following three major parts:

1. Low-complexity detection, based on a local neighborhood search and probabilistic data association (PDA), on large-MIMO links with channel state information at the receiver (CSIR) only, and the associated channel estimation.

2. Low-complexity precoding using X-Codes/X-Precoders and Y-Codes/Y-Precoders on large-MIMO links with channel state information at the transmitter (CSIT) and CSIR.

3. Low-complexity precoding for large multiuser MISO (multiple-input single-output) downlink systems with CSIT, based on vector perturbation with a reduced search space.

1. Low-Complexity Detection Using Local Neighborhood Search and PDA:
In this part of the work, we consider large-MIMO systems with channel state information at the receiver. We propose two low-complexity detection algorithms, one based...
on a local neighborhood search, termed as multistage likelihood ascent search ($M$-LAS) algorithm, and another based on probabilistic data association. We were motivated to investigate such algorithms from machine learning/artificial intelligence for the purpose of large-MIMO detection due to their demonstrated success in large systems including, for example, multiuser detection in code division multiple access (CDMA) systems with large number of users. These algorithms exhibit ‘large-system behavior,’ where the bit error rate (BER) performance improves and gets increasingly closer to the optimal performance for increasing number of antennas. We demonstrate the feasibility of these algorithms in both V-BLAST MIMO systems (which offer full rate) as well as non-orthogonal space-time block code (STBC) MIMO systems (which offer both full rate as well as full transmit diversity). The order of complexity for $M$-LAS algorithm is $O(N_t N_r)$ per symbol in V-BLAST MIMO, where $N_t$ and $N_r$ are the number transmit and receive antennas, respectively. We also propose a low-complexity iterative detection/channel estimation scheme. With the feasibility of such low-complexity detection/channel estimation schemes, large-MIMO systems with tens of antennas operating at several tens to hundreds of bps/Hz spectral efficiencies can become practical, enabling interesting high data rate wireless applications.

2. Low-Complexity Precoding Using X-Codes and Y-Codes:

In this part of the work, we consider a MIMO system with channel state information at both the transmitter and receiver. We propose X-Codes and Y-Codes to achieve high multiplexing and diversity gains at low complexity. The proposed precoding schemes are based upon the singular value decomposition (SVD) of the channel matrix which transforms the MIMO channel into parallel subchannels. Then X- and Y-Codes are used to improve the diversity gain by pairing the subchannels, prior to SVD precoding. In particular, subchannels with good diversity are paired with those having low diversity gains. Hence, a pair of channels is jointly encoded using a $2 \times 2$ real matrix, which is fixed a priori and does not change with each channel realization. For X-Codes, these matrices are 2-dimensional rotation matrices parameterized by a single angle, while for Y-Codes, these matrices are 2-dimensional upper left triangular matrices. The maximum likelihood (ML) decoding complexity for both X- and Y-Codes is low. Specifically, the decoding complexity of Y-Codes is the same as that of a scalar channel.
We also propose X-, Y-Precoders with the same structure as X-, Y-Codes, but the encoding matrices adapt to each channel realization. The optimal encoding matrices for X-, Y-Codes/Precoders are derived analytically. It is observed that X-Codes/Precoders perform better for well-conditioned channels, while Y-Codes/Precoders perform better for ill-conditioned channels, compared to other precoding schemes in the literature.

We then propose a non-diagonal precoder based on the X-Codes to increase the mutual information in Gaussian MIMO channels with discrete input alphabets. This precoding structure enables us to express the total mutual information as a sum of the mutual information of all the pairs. The problem of finding the optimal precoder with the above structure, which maximizes the total mutual information, is solved by i) optimizing the rotation angle and the power allocation within each pair and ii) finding the optimal pairing and power allocation among the pairs. It is shown that the mutual information achieved with the proposed pairing scheme is very close to that achieved with the optimal precoder by Cruz et al., and is significantly better than Mercury/waterfilling strategy by Lozano et al.

3. Low-Complexity Multiuser Precoding Using Reduced Search Space Vector Perturbation:

In this part of the work, we consider the problem of precoding in large multiuser MISO (multiple-input single-output) systems with large number of transmit antennas ($N_t$) at the base station and large number of downlink users ($N_u$), where each user has one receive antenna. Such large MISO systems are of interest because of the high capacities (sum-rates) of the order of tens to hundreds of bits/channel use possible in such systems. We propose a vector perturbation based low-complexity precoder, termed as norm descent search (NDS) precoder, which has a complexity of just $O(N_uN_t)$ per information symbol. This low complexity attribute of the precoder is achieved by searching for the perturbation vector over a reduced search space. Interestingly, in terms of BER performance, the proposed precoder achieves increasingly better BER for increasing $N_t, N_u$, making it suited for large MISO systems both in terms of complexity as well as performance.
## Glossary

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3GPP</td>
<td>Third Generation Partnership Project</td>
</tr>
<tr>
<td>ASIC</td>
<td>Application Specific Integrated Circuit</td>
</tr>
<tr>
<td>AWGN</td>
<td>Additive White Gaussian Noise</td>
</tr>
<tr>
<td>BC</td>
<td>BroadCast</td>
</tr>
<tr>
<td>BER</td>
<td>Bit Error Rate</td>
</tr>
<tr>
<td>BPCU</td>
<td>Bits Per Channel Use</td>
</tr>
<tr>
<td>BPSK</td>
<td>Binary Phase Shift Keying</td>
</tr>
<tr>
<td>BS</td>
<td>Base Station</td>
</tr>
<tr>
<td>CDA</td>
<td>Cyclic Division Algebra</td>
</tr>
<tr>
<td>CDMA</td>
<td>Code Division Multiple Access</td>
</tr>
<tr>
<td>CEP/CER</td>
<td>Codeword Error Probability/Rate</td>
</tr>
<tr>
<td>CI</td>
<td>Channel Inversion</td>
</tr>
<tr>
<td>CPE</td>
<td>Customer Premises Equipment</td>
</tr>
<tr>
<td>CPU</td>
<td>Central Processing Unit</td>
</tr>
<tr>
<td>CSI</td>
<td>Channel State Information</td>
</tr>
<tr>
<td>CSIR</td>
<td>CSI at the Receiver</td>
</tr>
<tr>
<td>CSIT</td>
<td>CSI at the Transmitter</td>
</tr>
<tr>
<td>DMG</td>
<td>Diversity-Multiplexing Gain</td>
</tr>
<tr>
<td>DPC</td>
<td>Dirty Paper Coding</td>
</tr>
<tr>
<td>DSL</td>
<td>Digital Subscriber Line</td>
</tr>
<tr>
<td>EE</td>
<td>Equal Energy</td>
</tr>
<tr>
<td>FD</td>
<td>Full Diversity</td>
</tr>
<tr>
<td>FPGA</td>
<td>Field Programmable Gate Arrays</td>
</tr>
<tr>
<td>HDTV</td>
<td>High-Definition TeleVision</td>
</tr>
<tr>
<td>IC</td>
<td>Interference Cancellation</td>
</tr>
<tr>
<td>IF</td>
<td>Intermediate Frequency</td>
</tr>
<tr>
<td>i.i.d.</td>
<td>Independent and Identically Distributed</td>
</tr>
<tr>
<td>ILL</td>
<td>Information LossLess</td>
</tr>
<tr>
<td>IPTV</td>
<td>Internet Protocol TeleVision</td>
</tr>
<tr>
<td>ISI</td>
<td>Inter-Symbol Interference</td>
</tr>
<tr>
<td>ISIC</td>
<td>Iterative Soft Interference Cancellation</td>
</tr>
<tr>
<td>LAS</td>
<td>Likelihood Ascent Search</td>
</tr>
<tr>
<td>LD</td>
<td>Linear Dispersion</td>
</tr>
<tr>
<td>LLR</td>
<td>Log-Likelihood Ratio</td>
</tr>
<tr>
<td>LOS</td>
<td>Line-Of-Sight</td>
</tr>
</tbody>
</table>
LTE-A : Long Term Evolution - Advanced
MAC : Media Access Control
MAP : Maximum a Posteriori
MF : Matched Filter
MIMO : Multiple-Input Multiple-Output
MISO : Multiple-Input Single-Output
ML : Maximum Likelihood
MLD : Maximum Likelihood Detection
M-LAS : Multistage LAS
MMSE : Minimum Mean Square Error
NDS : Norm-Descent Search
NLOS : Non-LOS
OFDM : Orthogonal Frequency Division Multiplexing
PAM : Pulse Amplitude Modulation
PDA : Probabilistic Data Association
QAM : Quadrature Amplitude Modulation
QOSTBC : Quasi-Orthogonal STBC
QPSK : Quadrature Phase Shift Keying
RF : Radio Frequency
SD : Sphere Decoder
SE : Sphere Encoder
SER : Symbol Error Rate
SIC : Successive Interference Cancellation
SIMO : Single-Input Multiple-Output
SINR : Signal-to-Interference plus Noise Ratio
SISO : Single-Input Single-Output
SNR : Signal-to-Noise Ratio
STBC : Space-Time Block Code
SVD : Singular Value Decomposition
TDD : Time Division Duplexing
THP : Tomlinson-Harashima Precoding
UWB : Ultra WideBand
V-BLAST : Vertical Bell Lab Space-Time architecture
VP : Vector Perturbation
WEP : Word Error Probability
WiMAX : Worldwide Interoperability for Microwave Access
WLAN : Wireless Local Area Network
WPAN : Wireless Personal Area Network
ZF : Zero Forcing
## Notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_t$</td>
<td>Number of transmit antennas</td>
</tr>
<tr>
<td>$N_r$</td>
<td>Number of receive antennas</td>
</tr>
<tr>
<td>$N_u$</td>
<td>Number of downlink users</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Average received SNR per receive antenna</td>
</tr>
</tbody>
</table>

**Boldface lower case letters**: Vectors

**Boldface upper case letters**: Matrices

- $j$: $\sqrt{-1}$
- $\Re(\cdot)$: Real part of the complex argument
- $\Im(\cdot)$: Imaginary part of the complex argument
- $(\cdot)^T$: Transposition
- $(\cdot)^H$: Hermitian transposition
- $(\cdot)^*$: Complex conjugation
- $E[\cdot]$: Expectation operator
- $\text{sgn}(\cdot)$: Signum function
- $\lfloor \cdot \rfloor$: Rounding operator
- $|\cdot|$: Absolute value of a complex number (or cardinality of a set)
- $\|\cdot\|$: Euclidean norm of a vector
- $\|\cdot\|_F$: Frobenius norm of a matrix
- $\text{tr}(\cdot)$: Trace of a matrix
- $\text{det}(\cdot)$: Determinant of a matrix
- $\text{vec}(\cdot)$: Stack columns of the input matrix into one column-vector

- $I_n$: $n \times n$ identity matrix
- $e_p$: Vector with its $p$th entry only as one and all other entries as zero
- $\lfloor c \rfloor$: Largest integer less than $c$
- $\mathbb{C}$: Field of complex numbers
- $\mathbb{R}$: Field of real numbers
- $\mathbb{R}^+$: Set of non-negative real numbers
- $\mathbb{Z}$: Ring of integers
- $\mathcal{CN}(\mu, \sigma^2)$: Circularly symmetric complex Gaussian distribution with mean $\mu$ and $\sigma^2$ variance
4 Large-MIMO Detection Using Probabilistic Data Association 96
   4.1 System Model ....................................................... 96
   4.2 Proposed PDA Based Detection ................................. 97
       4.2.1 Iterative Procedure ........................................ 98
       4.2.2 Complexity Reduction ..................................... 101
   4.3 Results and Discussions ........................................ 104
       4.3.1 BER Performance in Large V-BLAST MIMO ............... 104
       4.3.2 BER Performance in Large STBC MIMO ................. 105

5 Large-MIMO Precoding Using X- and Y-Codes 110
   5.1 System Model and SVD Precoding ............................. 111
   5.2 Pairing Good and Bad Subchannels ........................... 113
       5.2.1 ML Decoding .................................................. 115
       5.2.2 Performance Analysis ..................................... 117
       5.2.3 Design of Optimal Pairing ................................. 119
   5.3 X-Codes and X-Precoder ........................................ 120
       5.3.1 X-Codes and X-Precoders: Encoding ..................... 120
       5.3.2 X-Codes and X-Precoders: ML Decoding ............... 121
       5.3.3 Optimal Design of X-Codes ............................... 121
       5.3.4 Optimal Design of X-Precoder ............................ 125
   5.4 Y-Codes and Y-Precoder ........................................ 127
       5.4.1 Motivation .................................................... 127
       5.4.2 Y-Codes and Y-Precoders: Encoding ..................... 128
       5.4.3 Y-Codes and Y-Precoders: ML Decoding ............... 129
       5.4.4 Optimal Design of Y-Codes ............................... 131
       5.4.5 Optimal Design of Y-Precoder ............................ 133
   5.5 Simulation Results .............................................. 135
       5.5.1 Effect of Channel Condition on Error Performance .... 135
       5.5.2 Diversity Order Comparison ............................... 137
       5.5.3 Comparison of BER Performance with Full-rate Transmission .. 137
       5.5.4 Comparison of BER Performance for $N_t = N_r = 2, 4$ .... 138
       5.5.5 Comparison of BER Performance for $N_t = N_r = 16$ .... 140
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>F Proof of Theorem 3, Chapter 5</td>
<td>214</td>
</tr>
<tr>
<td>List of Publications from this Thesis</td>
<td>216</td>
</tr>
<tr>
<td>Bibliography</td>
<td>218</td>
</tr>
</tbody>
</table>
List of Figures

1.1 MIMO System. ................................................................. 4
1.2 Ergodic MIMO capacity for increasing $N_t = N_r$ with $i$) CSIR only and
    $ii$) CSIT and CSIR. ......................................................... 8
1.3 16-antenna indoor channel sounding measurement setup in [47]. Photos
    Source: [47]. ................................................................. 16
1.4 $64 \times 64$ MIMO indoor channel sounding setup at 5 GHz reported in [48].
    Photos Source: [48]. ......................................................... 17

2.1 Mean number of stages for the 3-LAS algorithm with MMSE initial vec-
    tor. V-BLAST MIMO, 4-QAM, SNR = 6, 9, 12 dB. ......................... 44
2.2 Mean number of iterations per stage per transmit antenna for the 3-LAS
    algorithm with MMSE initial vector. V-BLAST MIMO, 4-QAM, SNR =
    6, 9, 12 dB. ................................................................. 45
2.3 Computational complexity of the $M$-LAS algorithm in decoding non-
    orthogonal STBCs from CDA. MMSE initial vector, 4-QAM, SNR = 6 dB. 47
2.4 Uncoded BER performance of 3-LAS detection for different values of
    $N_t = N_r = (16, 32, 64, 128, 256)$ and 4-QAM. MMSE initial filter. .... 50
2.5 Average received SNR required to achieve a target BER of $10^{-3}$ in V-
    BLAST MIMO for increasing values of $N_t = N_r$ using 1-LAS detection
    with MMSE initial vector. 4-QAM. ......................................... 51
2.6 Average received SNR required to achieve a target BER of $10^{-4}$ in V-
    BLAST MIMO for increasing values of $N_t = N_r$ using 1-LAS detection
    with MMSE initial vector. 16-QAM. ........................................ 51
2.7 Comparison of 3-LAS and 1-LAS performance in V-BLAST MIMO for $N_t = N_r = 64, 32$ and 4-QAM. MMSE initial vector. ............................... 52
2.8 Turbo coded BER performance of 3-LAS detection for $N_t = N_r = 64, 128$, 4-QAM, and rate-3/4 turbo code. MMSE initial vector. ............................. 53
2.9 Uncoded BER performance of 1-LAS, 2-LAS and 3-LAS detectors for ILL-only STBCs for different $N_t = N_r$. 4-QAM, $2N_t$ bps/Hz. ............... 54
2.10 Uncoded BER performance comparison between FD-ILL and ILL-only STBCs for different $N_t = N_r$. 4-QAM, $2N_t$ bps/Hz, 1-LAS detection with MMSE initial vector. ................................................................. 55
2.11 Uncoded BER comparison between perfect codes and ILL-only STBCs for different $N_t = N_r$, 4-QAM, $2N_t$ bps/Hz, 1-LAS detection. .......... 57
2.12 Uncoded BER performance comparison between perfect codes, ILL-only, and FD-ILL STBCs for $N_t = N_r = 16, 32$, 16-QAM, $4N_t$ bps/Hz, 1-LAS detection. ................................................................. 58
2.13 Uncoded BER performance comparison between the 1-LAS algorithm and the ISIC algorithm in [111] for ILL-only STBCs for different $N_t = N_r$, 4-QAM, $2N_t$ bps/Hz. MMSE initial vectors for both 1-LAS and ISIC. ... 59
2.14 Uncoded BER comparison between different large-MIMO architecture/detector combinations for given number of transmit/receive antennas ($N_t = N_r = 16$) and spectral efficiency (32 bps/Hz). ............................... 60
2.15 Turbo coded BER performance of 1-LAS detector for $16 \times 16$ FD-ILL and ILL-only STBCs. $N_t = N_r = 16$, 4-QAM, turbo code rates: 1/3, 1/2, 3/4 (10.6, 16, 24 bps/Hz). ............................... 63
2.16 Turbo coded BER performance of 1-LAS detector for $32 \times 32$ ILL-only STBC. $N_t = N_r = 32$, 16-QAM, turbo code rates: 1/3, 1/2, 3/4 (42.6, 64, 96 bps/Hz). ............................... 64
2.17 Turbo codeword error probability of 1-LAS detection in a slow fading MIMO channel for $N_t = N_r = 4, 8, 12$, 4-QAM, rate-3/4 turbo code and FD-ILL STBCs. MMSE initial vector. ............................... 64
2.18 Propagation scenario for the MIMO fading channel model. ............................... 67
2.19 Effect of $N_r > N_t$ in correlated MIMO fading channel model in [8] keeping $N_rd_r$ constant and $d_t = d_r$. $N_rd_r = 72$ cm, $f_c = 5$ GHz, $R = 500$ m, $S = 30$, $D_t = D_r = 20$ m, $\theta_t = \theta_r = 90^\circ$, 12 × 12 ILL-only STBC, $N_t = 12$, $N_r = 12, 18$, 16-QAM, rate-3/4 turbo code, 36 bps/Hz. .......................... 68

2.20 Transmission scheme with one pilot matrix followed by $N_d$ data STBC matrices in each frame. ................................................................. 71

2.21 Hassibi-Hochwald (H-H) capacity bound for 1P+8D ($T = 144$, $\tau = 16$, $\beta_p = \beta_d = 1$) and 1P+1D ($T = 32$, $\tau = 16$, $\beta_p = \beta_d = 1$) training for a 16 × 16 MIMO channel. Perfect CSIR capacity is also shown. ......................... 73

2.22 Uncoded BER of 1-LAS detector for 16 × 16 ILL-only STBC with i) perfect CSIR, ii) CSIR using MMSE estimation scheme, and iii) CSIR using iterative detection/channel estimation scheme (4 iterations). $N_t = N_r = 16$, 4-QAM, 1P+1D ($T = 32$, $\tau = 16$, $\beta_p = \beta_d = 1$) and 1P+8D ($T = 144$, $\tau = 16$, $\beta_p = \beta_d = 1$) training. ....................................................... 74

2.23 Turbo coded BER performance of 1-LAS detector for 16 × 16 ILL-only STBC with i) perfect CSIR, ii) CSIR using MMSE estimation, and iii) CSIR using iterative detection/channel estimation (4 iterations). $N_t = N_r = 16$, 4-QAM, rate-3/4 turbo code, 1P+1D ($T = 32$, $\tau = 16$, $\beta_p = \beta_d = 1$) and 1P+8D ($T = 144$, $\tau = 16$, $\beta_p = \beta_d = 1$) training. ............................. 75

2.24 Turbo coded BER performance of 1-LAS detection and iterative estimation/detection as a function of coherence time, $T = 32, 144, 400, 784$, for a given $N_t = N_r = 16$, 16 × 16 ILL-only STBC, 4-QAM, rate-3/4 turbo code. ................................................. 76

2.25 Comparison between two 1P+$N_d$D training-based systems, one with a larger $N_t$ than the other for a given $N_r$ and $T$. ................................. 77

3.1 Simulated pdf of $z_{u_n,d}$ for $n = 2N_t$ for increasing $N_t = N_r$. 4-QAM. The pdf tends towards Dirac delta function at zero. ................... 85

3.2 Conditional probabilities $p_{r, r+1}$ as a function of increasing $N_t = N_r$ for the 1-LAS detector with MMSE initial vector. 4-QAM and SNR = 10 dB. 87

3.3 Conditional probabilities $p_{r, r+1}$ as a function of increasing $N_t = N_r$ for the 1-LAS detector with Random initial vector. 4-QAM and SNR = 10 dB. 88
3.4 Simulated BER performance of the 1-LAS detector for V-BLAST MIMO as a function of average received SNR for increasing values of $N_t = N_r$. MMSE initial vector, 4-QAM. .......................................................... 92

3.5 Simulated BER performance of 1-LAS detector with Random initial vector for increasing $N_t = N_r$. 4-QAM. .......................................................... 93

3.6 Simulated BER performance of the 1-LAS detector with MF initial vector for increasing $N_t = N_r$. 4-QAM. .......................................................... 93

3.7 BER performance comparison between Random 1-LAS and Random 3-LAS detectors for increasing $N_t = N_r$. 4-QAM. .......................................................... 94

4.1 BER performance of PDA based detection in a V-BLAST MIMO system with $N_t = N_r = 64$ and 4-QAM for different number of PDA iterations ($m = 1, 2, 4, 8$). .......................................................... 104

4.2 BER performance of PDA based detection of V-BLAST MIMO for $N_t = N_r = 8, 16, 32, 64, 96$, 4-QAM, and $m = 5$ iterations. .......................................................... 105

4.3 Comparison between the uncoded BER performance of PDA and LAS algorithms in decoding $4 \times 4, 8 \times 8, 16 \times 16$ ILL-only STBCs. $N_t = N_r$, 4-QAM, $m = 10$. .......................................................... 106

4.4 Comparison between the uncoded BER performance of PDA and LAS algorithms in decoding $4 \times 4, 8 \times 8, 16 \times 16$ ILL-only STBCs. $N_t = N_r$, 16-QAM. .......................................................... 107

4.5 Turbo coded BER performance of the PDA algorithm in decoding $12 \times 12$ ILL-only STBC with $N_t = N_r = 12$, 4-QAM, rate-$3/4$ turbo code, 18 bps/Hz and $m = 10$ for (i) perfect CSIR, and (ii) estimated CSIR using 2 iterations between PDA decoding/channel estimation. .......................................................... 107

4.6 Effect of spatial correlation on the performance of PDA in decoding $12 \times 12$ ILL STBC from CDA. $N_t = 12$, $N_r = 12, 18$, 16-QAM, rate-$3/4$ turbo code, 36 bps/Hz. Correlated channel parameters: $f_c = 5$ GHz, $R = 500$ m, $S = 30$, $D_t = D_r = 20$ m, $\theta_t = \theta_r = 90^\circ$, $N_r d_r = 72$ cm, $d_t = d_r$. .......................................................... 109

5.1 Union bound for the word error probability. $N_t = N_r = 2$ and $M = 2$ (4-QAM). .......................................................... 123
5.2 Sensitivity of word error probability w.r.t. $\theta_1$. $N_r = N_t = 2$ and $M = 2, 4$ (4-QAM, 16-QAM). .......................................................... 124
5.3 One quadrant of the set $S_M$ for $M = 2, 4$ (4-QAM, 16-QAM). The critical angles where performance degrades severely are shown to coincide with $\tan^{-1}(-p/q)$. .......................................................... 125
5.4 Received signal space for the real component of the $k$th pair. $M = 8$ and so we have 5 regions with vertical dashed lines demarcating the boundary between the regions. The scaled codebook vectors are represented by small filled circles along with their corresponding codebook index number. Dotted lines demarcate the boundary between the ML decision regions. .......................................................... 131
5.5 Effect of the channel condition number on error performance of various precoders for a $2 \times 2$ system with target spectral efficiency equal to 8 bps/Hz. .......................................................... 136
5.6 BER comparison between various precoders for $N_t = N_r = N_s = 2, 4$ and $M = 2$ (4-QAM). Target spectral efficiency is equal to $2N_s$ bps/Hz. 138
5.7 BER comparison between various precoders for $N_t = N_r = 2$ and target spectral efficiency = 4, 8 bps/Hz. .......................................................... 139
5.8 BER comparison between various precoders for $N_t = N_r = 4$ and target spectral efficiency = 8, 16 bps/Hz. .......................................................... 139
5.9 BER comparison between various precoders for $N_t = N_r = 16$ and target spectral efficiency = 64 bps/Hz. .......................................................... 141
5.10 BER comparison between the proposed X-Codes and Y-Codes for $N_t = N_r = 2$ with spectral efficiency = 4, 8 bps/Hz. .......................................................... 142
5.11 BER comparison between the proposed X-Codes and Y-Codes for $N_t = N_r = 4$ with spectral efficiency = 8, 16 bps/Hz. .......................................................... 142
5.12 Word error probability comparison between the proposed suboptimal Y-Precoders and exact optimal Y-Precoders for $N_t = N_r = 2, 4$ with spectral efficiency = $4N_t$ bps/Hz. .......................................................... 144
6.1 Plot of $f^{opt}$ versus $P_T$ for $n = 2$ parallel channels with $\beta = 1, 1.5, 2, 4, 8$ and $\alpha = 1$. Input alphabet is 16-QAM. .......................................................... 158
6.2 Plot of $\theta_{opt}$ versus $P_T$ for $n = 2$ parallel channels with $\beta = 1.5, 2, 4, 8$ and $\alpha = 1$. Input alphabet is 16-QAM. ................................................................. 158

6.3 Mutual Information of X-Codes versus power allocation fraction $f$ for $n = 2$ parallel channels with $\beta = 1, 1.5, 2, 4, 8$, $\alpha = 1$ and $P_T = 17$ dB. Input alphabet is 16-QAM. ................................................................. 159

6.4 Mutual information of X-Codes versus rotation angle $\theta$ for $n = 2$ parallel channels with $\beta = 1, 1.5, 2, 4, 8$, $\alpha = 1$ and $P_T = 17$ dB. Input alphabet is 16-QAM. ................................................................. 160

6.5 Mutual information versus $P_T$ for X-Codes for different $\theta$'s, $n = 2$ parallel channels, $\alpha = 1$, $\beta = 2$, and 4-QAM input alphabet. ........................................ 160

6.6 Mutual information versus $P_T$ for $n = 2$ parallel channels with $\beta = 2$ and $\alpha = 1$, for 4-QAM and 16-QAM. ................................................................. 161

6.7 Mutual information versus $P_T$ for $n = 2$ parallel channels with varying $\beta = 1, 2, 4$, $\alpha = 1$ and 4-QAM input alphabet. ................................................................. 162

6.8 Mutual information with X-Codes versus $P_T$ for $n = 2$ parallel channels with varying $\beta = 1, 1.5, 2, 4, 8$, $\alpha = 1$ and 16-QAM input alphabet. .... 164

6.9 Mutual information versus $P_T$ with two different pairings for a $n = 4$ diagonal channel and 16-QAM input alphabet. ................................................................. 165

6.10 Mutual information versus $P_T$ for the Gigabit DSL channel given by (42) in [95]. ................................................................. 165

6.11 Comparison of $4 \times 4$ wireless MIMO ergodic capacity with perfect CSIT/CSIR for 16-QAM input alphabet. ................................................................. 166

6.12 Mutual information versus per subcarrier SNR for an OFDM system with 32 carriers. X-Codes versus Mercury/waterfilling. ....................... 169

6.13 Mutual information versus per subcarrier SNR for a OFDM system with 32 carriers. Comparison of heuristic pairing schemes. ............ 170

6.14 Comparison of $16 \times 16$ wireless MIMO ergodic capacity with perfect CSIT/CSIR for 16-QAM input alphabet. ....................... 172

7.1 Multiuser MISO system on the downlink. ........................................ 174

7.2 Proposed norm-descent search precoder. ........................................ 179
7.3 Average number of iterations per information symbol till the NDS algorithm terminates as a function of \((N_t, N_u)\). \(N_t = N_u, N_r = 1, 4\)-QAM, \(G_{\text{MMSE}}\) precoding matrix. .................................................. 183

7.4 Uncoded BER performance of the proposed NDS-MMSE and NDS-ZF precoders for \(N_t = N_u = 10, 50, 200. N_r = 1, 4\)-QAM. ................................. 184

7.5 Uncoded BER performance comparison of the proposed NDS-MMSE precoder versus i) MMSE-only precoder, and ii) the VP-SE scheme. \(N_u = 8\) and \(N_t = 8, 16, N_r = 1, 4\)-QAM. ........................................ 185

7.6 Complexity of the proposed NDS-MMSE precoder in comparison with that of the VP-SE scheme. \(N_t = N_u, N_r = 1, 4\)-QAM, SNR = 15 dB. ........ 187

7.7 Upper and lower bounds for the ergodic sum capacity, \(C_{\text{sum}}\). ........ 189

7.8 Turbo coded BER performance comparison between the NDS-MMSE and VP-SE schemes for \(N_u = 8, N_t = 8, 12, 16, N_r = 1, 4\)-QAM, rate-3/4 turbo code, sum rate = 12 bps/Hz. ................................. 190

7.9 Turbo coded BER performance of the proposed NDS-MMSE precoder without and with channel estimation errors. \(N_t = N_u = 300, N_r = 1, 4\)-QAM, rate-3/4 turbo code, sum rate = 450 bps/Hz. ........ 191
List of Tables

2.1 Complexity and performance comparison of different large-MIMO architecture/detector combinations, all with $N_t = N_r = 16$ and achieving 32 bps/Hz spectral efficiency. .................................................. 61

2.2 On optimum $N_t$ for a given $N_r$ and $T$. System-II with a smaller $N_t$ achieves a higher spectral efficiency while achieving $10^{-3}$ coded BER at a lesser SNR than System-I with a larger $N_t$. .......................... 78

4.1 Listing of the proposed PDA based detection algorithm. ................. 103

7.1 Listing of the proposed NDS precoding algorithm. ....................... 182
Chapter 1

Introduction

Multiple-input multiple-output (MIMO) systems with multiple antennas at the transmitter and receiver sides have become very popular owing to the several advantages they promise to offer, including high data rates and transmit diversity [1]-[4]. For example, current wireless standards including WiFi (IEEE 802.11n) and WiMAX (IEEE 802.16e) and 3GPP LTE-A have adopted MIMO techniques in their physical layer. It is known that the capacity of MIMO channels grows linearly with the minimum of the number of antennas on the transmitter and receiver sides [5],[6], which motivates the use of large number of antennas at the communication terminals to achieve high spectral efficiencies as well as high diversity orders in wireless transmissions.

1.1 Multi-Antenna Wireless Channels

Multi-antenna wireless channels are a broad category of channels that include point-to-point and multiuser channels (e.g., MAC, broadcast, and relay channels, to name a few). One of the defining characteristics of a wireless channel is the variation of the channel strength over time and over frequency [4],[7]. These variations are typically classified into two types: large-scale fading and small-scale fading. Large-scale fading is due to path loss of the signal as a function of distance and shadowing by
large objects like buildings, bridges and trees, and is typically frequency independent. Small-scale fading, on the other hand, is due to the constructive and destructive interference of the multiple signal paths between the transmitter and receiver. Small-scale fading happens at the spatial scale of the order of the carrier wavelength, and is frequency dependent which makes the channel to be classified as frequency-selective or frequency-flat. When the signaling bandwidth is larger than the coherence bandwidth of the channel (which has an inverse relation with the maximum delay spread of the channel), the channel is frequency-selective. In frequency-flat channels, the signaling bandwidth is much smaller than the coherence bandwidth of the channel [7]. In this thesis, we will consider the channel to be frequency-flat. Even when the channel is frequency-selective, techniques like orthogonal frequency division multiplexing (OFDM) can convert the channel into multiple frequency-flat channels on which the techniques proposed in this thesis can be employed. In terms of time variation, wireless channels are further classified as slowly fading or fast fading, depending on the fade rate relative to the signaling rate. If the fade remains constant over the signaling duration, the fading is termed as slow (or time-flat) fading, whereas if the fade varies within the signaling duration, it is termed as fast (or time-selective) fading. Carrier wavelength and velocity of the communication terminal determine the amount of time-selectivity (or Doppler spread) in the channel [7].

Most multi-antenna wireless channels with $N_r$ receive and $N_t$ transmit antennas (e.g., see Fig. 1.1) are modeled as a linear channel with an equivalent baseband channel matrix $H_c \in \mathbb{C}^{N_r \times N_t}$. The $(i, j)$th entry of $H_c$, i.e., $h_{i,j}$ represents the channel gain from the $j$th transmit antenna to the $i$th receive antenna. These channel gains can be independent or correlated, which depends upon various factors, including spacing between transmit and receive antennas, amount of scattering in the environment, pin-hole effects, etc. [8]. Mathematical models that characterize the spatial correlation in MIMO channels are used in the performance evaluation of MIMO systems. Spatial correlation
at the transmit and/or receive side can affect the rank structure of the MIMO channel resulting in degraded MIMO capacity [9]. The structure of scattering in the propagation environment can also affect the capacity [8]. Also, transmit correlation in MIMO fading can be exploited by using non-isotropic inputs (precoding) based on the knowledge of the channel correlation matrices [10]-[12].

The channel gains are also referred to as the channel state information (CSI). Availability of the knowledge of these gains at the receiver and transmitter is an important factor which decides the performance of the communication system. CSI at the receiver (CSIR) refers to the scenario where the receiver has the knowledge of the channel gains. Likewise, CSI at the transmitter (CSIT) refers to the scenario where the transmitter has the knowledge of the channel gains. In fast fading channels, accurate estimation of the channel gains can become an issue, in which case non-coherent or blind techniques can be considered. In addition, feedback based provision of CSIT can become ineffective in fast fading. However, in applications where the channel is not varying fast, it is generally possible to estimate the channel gains accurately through pilot-assisted transmission. Also, CSIT based on feedback from measured CSI from the receiver is effective in such slow fading channels.

### 1.2 MIMO System Model

Assuming frequency-flat and slow fading, where the channel gains are assumed to remain constant over the signaling interval, the equivalent complex baseband MIMO system model can be written as

\[
y_c = H_c x_c + n_c,
\]

where \( x_c \in \mathbb{C}^{N_t} \) is the transmitted vector, \( y_c \in \mathbb{C}^{N_r} \) is the received vector, and \( n_c \in \mathbb{C}^{N_r} \) is the additive white Gaussian noise (AWGN) vector. The \( j \)th entry of \( x_c \) is the symbol.
transmitted from the $j$th transmit antenna, $j = 1, \cdots, N_t$. In a typical communication system, information bits (e.g., output of some source coder, like image or voice compression, followed by a channel coder) are grouped into messages, and each message then corresponds to an $N_t$-dimensional complex vector, whose $j$th component is transmitted from the $j$th transmit antenna. In practice, these vectors belong to some codebook $\mathcal{X}$. The transmitter groups $R = \log_2 |\mathcal{X}|$ bits into a message, which is then used to index into the codebook. $R$ is often referred to as the rate of the codebook or simply as the rate of transmission. Alternatively, the $x_c$ vector could be a pilot symbol vector (known to the receiver) during the training phase in a pilot-aided channel estimation scheme. The $i$th entry of $y_c$ is the signal received at the $i$th receive antenna, $i = 1, \cdots, N_r$. Assuming a rich scattering environment, the entries of the channel matrix $H_c$ are often modeled as i.i.d. (independent and identically distributed) $\mathcal{CN}(0, 1)$.

Since the transmitter is power constrained, we have $E[\text{tr}(x_c x_c^H)] = P$, where $P$ is the total power available at the transmitter. Also, $E[n_c n_c^H] = \sigma^2 I_{N_r}$, where $\sigma^2$ is the noise variance at each receive antenna. The average received signal-to-noise ratio (SNR) at each receive antenna is given by $\gamma = \frac{P}{\sigma^2}$. 
The MIMO signal detection problem can be stated as: Given $y_c$ and the knowledge of $H_c$, determine $\hat{x}_c$, an estimate of the transmitted symbol vector $x_c$. Likewise, the MIMO channel estimation problem in a training-based scheme can be stated as: Given the knowledge of the transmitted pilot symbol vector $x_c$, determine $\hat{H}_c$, an estimate of the channel gain matrix $H_c$. A vast body of the MIMO literature has focused on these two problems [13]. Joint detection/channel estimation approaches have also been investigated [14],[15]. Much of the MIMO detection and channel estimation works in the literature so far, however, have considered only small number of antennas (e.g., $N_t = 2, 3, 4$). In the first part of this thesis (Chapters 2, 3, 4), we will be concerned with MIMO detection and channel estimation when $N_t, N_r$ are large (e.g., tens).

### 1.3 MIMO Communication with CSIR Only

In communication channels, error probability is one of the key performance indicators. Most communication schemes employ channel coding schemes to increase robustness against errors. To achieve an arbitrarily low probability of error, the rate of transmission $R$ must be strictly below the MIMO channel capacity. The MIMO channel capacity is dependent on $H_c$ and the transmit covariance matrix $K_x \triangleq \mathbb{E}[x_c x_c^H]$, $\text{tr}(K_x) = P$, and is given by

$$C_{\text{mimo}}(\gamma, H_c, K_x) = \log_2 \det \left( \mathbf{I}_{N_r} + \frac{1}{\sigma^2} H_c K_x H_c^H \right). \quad (1.2)$$

In the case of availability of CSIR only, since the transmitter has no knowledge of the channel gains, it cannot adapt its transmission scheme with respect to the channel gains. Therefore, for a fixed $\gamma$ and rate $R$, the transmitter uses a fixed codebook, which does not change with changing channel gains. The transmitter codebook selection is very much dependent on whether the channel is slow fading or fast fading.
1.3.1 Slow Fading Channels

In slowly fading channels, where the channel does not change during the length of the codeword, if the channel is such that \( C_{\text{mimo}}(\gamma, H_c, K_x) < R \), then no detector can recover the transmitted codeword correctly, and the channel is said to be in outage. Hence, for slow fading channels with CSIR-only, outage cannot be avoided and it is impossible to achieve an arbitrary low probability of error. In such scenarios, an appropriate performance indicator of any encoding-decoding scheme is the codeword error probability (CEP) or codeword error rate (CER). For codewords of large length, the theoretical limit for the codeword error rate of any encoding-decoding scheme is the channel outage probability, which is defined as

\[
P_{\text{outage}}(\gamma, R) = \min_{K_x | \text{tr}(K_x) = P} p(C_{\text{mimo}}(\gamma, H_c, K_x) < R). \quad (1.3)
\]

Any practical encoding-decoding scheme would have a codeword error rate more than the channel outage probability given in (1.3). Therefore, it is important to design transmit schemes and corresponding receivers which can perform very close to the channel outage probability for all values of \( \gamma \) and \( R \). For slowly fading channels, there are two important parameters, namely, diversity gain and multiplexing gain. Diversity gain is a measure of reliability, whereas multiplexing gain is a measure of the degrees of freedom in the MIMO channel. These two parameters are usually related by the so called diversity-multiplexing gain (DMG) tradeoff [16]. It is known that in Rayleigh faded MIMO channels, the maximum diversity gain achievable is \( N_r N_t \), and the maximum multiplexing gain achievable is \( \min(N_r, N_t) \). When the rate of transmission \( R \) is fixed, the limiting value (as \( \gamma \to \infty \)) of the negative of the slope of \( \log(P_{\text{outage}}(\gamma, R)) \) w.r.t. \( \log \gamma \) can be no more than \( N_r N_t \). For a given scheme, we can therefore define the diversity order achievable (with fixed \( R \)) as

\[
d_{\text{ord}} = - \lim_{\gamma \to \infty} \frac{\log(P_e(\gamma))}{\log \gamma}, \quad (1.4)
\]
where $P_e(\gamma)$ is the codeword error rate of the scheme. For simple MIMO schemes like V-BLAST [17],[18], it can be shown that the maximum diversity order achievable is only $N_r$. This is because symbols transmitted from the antennas are independent, and each such symbol reaches the receiver only through $N_r$ different paths.

Space-time block coding is a well known technique [2], which can achieve the full diversity gain of $N_r N_t$. To achieve full diversity, symbols are coded across both space and time. This implies that joint detection be performed at the receiver, which makes the receiver processing complex. Orthogonal space-time block codes (STBC) were proposed as a solution to this problem [19],[20]. They, however, sacrificed on the multiplexing rate, and were therefore not suited for systems with high target spectral efficiencies. Subsequent to orthogonal STBCs, several high-rate and high-diversity STBCs were proposed. One such class of STBCs is non-orthogonal STBCs from Cyclic Division Algebras (CDA) [21],[22]. STBCs from CDA can achieve the full diversity of $N_r N_t$ without sacrificing on the rate. However, due to coding across space and time, receiver processing is challenging. The maximum-likelihood (ML) detector has a complexity that is exponential in the number of symbols to be detected jointly. For e.g., the ML detector for even a $4 \times 4$ full-rate STBC from CDA with 16-QAM would have to check the ML metric for $16^{16}$ different STBC matrices. Hence, suboptimal detectors that have lesser complexity and can achieve near-optimal performance are of interest. Chapters 2, 3, and 4 in this thesis are concerned with low-complexity algorithms that can achieve near-optimal performance in V-BLAST MIMO as well as non-orthogonal STBC MIMO systems where $N_t, N_r$ are large (e.g., tens).

1.3.2 Fast Fading Channels

In fast fading channels, the channel fade changes multiple times during the duration of the codeword. By spreading portions of the codeword across multiple fades, reliability of codeword reception can be improved. In such a scenario, if the MIMO channel is
ergodic, in the limit of infinitely long codewords, it is possible to achieve error free communication if the rate of transmission $R$ satisfies

$$R \leq C_{\text{ergodic}}(\gamma) \triangleq \max_{K_x | H(K_x) = P} \mathbb{E}_{H_c} \left[ C_{\text{mimo}}(\gamma, H_c, K_x) \right].$$  \tag{1.5}$$

$C_{\text{ergodic}}(\gamma)$ is often referred to as the ergodic MIMO capacity, and is achieved with $K_x = \frac{P}{N_t} I_{N_t}$. This transmit architecture is also known as the V-BLAST scheme, where the symbol streams transmitted from each transmit antenna are uncorrelated. The ergodic MIMO capacity is therefore given by

$$C_{\text{ergodic}}(\gamma) = \mathbb{E}_{H_c} \left[ \log_2 \det \left( I_{N_r} + \frac{\gamma}{N_t} H_c H_c^H \right) \right].$$ \tag{1.6}$$

It has been shown that the ergodic MIMO capacity increases linearly with increasing $N_r = N_t$ [5].

In Figure 1.2, the ergodic MIMO capacity is plotted as a function of the average received SNR, $\gamma$, for MIMO systems with CSIR-only for different values of $N_t = N_r$. We observe that for a given SNR, the ergodic MIMO capacity increases linearly with
$N_t = N_r$. For e.g., at an SNR of 6.8 dB the ergodic capacity is 16, 32 and 64 bps/Hz for $N_t = N_r = 8, 16, 32$, respectively. This implies that at an SNR of 6.8 dB, a $N_t = N_r$ MIMO system would have an ergodic capacity of roughly $2N_t$ bps/Hz. Hence, MIMO systems with large $N_t = N_r$ (which we refer to as ‘large-MIMO systems’) have the potential of achieving very high spectral efficiencies of the order of tens to hundreds of bps/Hz. However, achieving such large spectral efficiencies requires that large number of independent spatial dimensions exist. This would depend on spacing between antennas, carrier wavelength, scattering around the wireless terminals, and spatial correlation. These requirements can be favorably met in several indoor/outdoor wireless applications with medium-/large-sized aperture communication terminals (e.g., set top boxes, laptops, TVs, fixed nodes in wireless mesh networks). We will discuss the technological challenges in realizing large-MIMO systems in Section 1.6.

1.4 MIMO Communication with CSIT and CSIR

If MIMO systems are operating in time division duplex (TDD) mode or if MIMO channels are slowly varying, it is possible for both the transmitter as well as the receiver to acquire the channel state information. When both CSIT and CSIR are available, the ergodic MIMO capacity is known to be achieved with independent Gaussian inputs beamformed along the right singular vectors of the channel matrix. This transforms the MIMO channel into a set of parallel non-interfering $n = \min(N_t, N_r)$ subchannels. Capacity is then achieved by waterfilling power allocation among these $n$ subchannels [5]. Note that the optimal power allocation was isotropic for the case of CSIR-only.

Due to the availability of CSIT, it is possible to use the available power judiciously by allocating more power to the subchannel with higher channel gain. At low SNR, availability of CSIT has an even higher impact on the ergodic capacity when compared to the CSIR-only scenario. This is because, at low SNR, capacity is known to increase
almost linearly with SNR, and therefore with CSIT, the transmitter allocates all available power to the subchannel with the highest channel gain. In contrast to this, with CSIR-only, the available power is equally divided among the subchannels, resulting in a lesser achievable capacity when compared to the CSIT scenario. At high SNR, waterfilling power allocation distributes roughly equal power to all the subchannels. Therefore, power allocation at high SNR is almost same for both the scenarios of CSIR-only and CSIT. This implies that at high SNR, both CSIT and CSIR-only scenarios have roughly the same ergodic capacity. These observations are illustrated through Fig. 1.2, where we plot the ‘CSIT and CSIR’ ergodic capacity, in addition to the ‘CSIR-only’ capacity. It is observed that indeed, for a given $N_t = N_r$ and SNR $\gamma$, the ergodic capacity with ‘CSIT and CSIR’ is more than the ergodic capacity with ‘CSIR-only’. Also, the gap between the ergodic capacity of the two scenarios reduces with increasing SNR.

Another important fact, which is not highlighted in Fig. 1.2 is that, at low SNR the ergodic capacity with ‘CSIT and CSIR’ is more than $n \log_2(1 + \gamma)$, which is the capacity of $n$ parallel and independent single-input single-output (SISO) non-faded AWGN channels [4].

In slowly fading channels, the codewords transmitted are subject to block fading (i.e., the channel remains almost same for the whole duration of the transmitted codeword). As pointed out earlier, in such block fading scenarios, if the capacity of the channel is below the rate of transmission, there will always be a codeword error (outage) irrespective of the coding scheme used. With the availability of CSIT, however, it is possible to theoretically achieve zero outage probability by adapting the transmitted codewords (i.e., codeword rate and transmit power) for a given long-term average power constraint [23]. This leads to a variable rate transmission scheme, and also large peak to average requirement on the transmit RF amplifiers, which are undesirable in many applications. Therefore, in such applications, it is obvious that outages cannot be avoided. Hence, it is important that encoding and decoding schemes be devised to achieve the
high diversity and multiplexing gains. Note that the maximum diversity gain is $N_t N_r$ and the maximum multiplexing gain is $\min(N_t, N_r)$. CSIT can be used to encode the information symbols into transmit vectors, a process commonly called as ‘precoding.’ Several precoding schemes are known in the literature [24]. Most known precoding schemes (or precoders in short) achieve either $i$) high-rate or high-diversity at low complexity (e.g., linear precoders like those proposed in [25],[26], and non-linear precoders [27] which are based on Tomlinson-Harashima precoding [28],[29]), or $ii$) both high-rate and high-diversity but at high complexity (e.g., precoders based on lattice reduction techniques [30] and vector perturbation [31],[32]). Motivated by this trade-off, in the second part of this thesis (Chapters 5, 6), we propose precoding schemes, termed as X-Codes/X-Precoders and Y-Codes/Y-Precoders, which can achieve both full-rate AND high-diversity at low complexity. We also show that the proposed X-Codes based precoding can be used to maximize the mutual information for discrete input alphabets, and achieve close to optimal precoding capacity with discrete inputs at low complexities.

1.5 Multiuser MIMO Communication

There is growing interest in MIMO techniques applied to multiuser communications, which is a broad topic in advanced communication theory [4],[33]. There are various types of communication models; e.g., broadcast (BC), multiple access (MAC), relaying, cooperation, etc. Of particular interest is downlink communications where a base station (BS) equipped with multiple transmit antennas sends data to multiple downlink users, each having one receive antenna [34].

In a broadcast channel, a transmitter (typically, a BS) sends data which contains information for all users in the system. The users extract out information pertinent to them. In such multiuser communication scenarios, the set of all rates achievable by the users is called the rate region. Given a multiuser channel (which comprises of the individual
channel gains from the transmitter to each user), there exists a corresponding rate region. The rate region for the Gaussian BC has been characterized, and it is known that a subset of this region can be achieved by dirty-paper coding (DPC) \[35\],\[36\]. The sum capacity (i.e., maximum aggregation of all users’ data rates, that grows linearly with the minimum of the number of antennas, $N_t$, and the number of users $N_u$, provided the transmitter and receivers all know the channel) was shown to be achievable by DPC \[37\],\[38\]. In DPC, the information is coded is such a way that, despite interference from other users, each user can still receive its information perfectly as if there were no interference at all. DPC is, however, not suited to practical implementations due to high complexity. Hence, practical encoding and decoding schemes are required to achieve the rates in the BC rate region.

Practical MIMO precoding that aims to achieve the sum capacity promised by DPC have been a key research topic \[28\]-\[32\],\[39\]-\[45\]. A simple idea that seems to give reasonable performance is the idea of pre-nulling the interference from users by precoding the information vector with the inverse of the channel matrix. This precoder (also known as the zero-forcing precoder \[39\]) does indeed get rid of interference, but the penalty to be paid is in the increase of average transmit power (particularly when the channel is ill-conditioned). Due to this, the zero-forcing precoder is known to achieve poor diversity order in fading channels. Most other low complexity precoders also suffer from loss in performance.

Vector perturbation (VP) based techniques were proposed as a low complexity alternative to DPC \[31\],\[32\]. VP techniques are known to achieve good performance. In VP based techniques, the precoder matrix is still the channel inverse matrix. However, prior to zero-forcing, the information symbol vector is perturbed by an integer vector, in such a way that the transmit power requirement is minimized. The optimal integer vector is usually searched using sphere encoding or lattice reduction techniques. Though the complexity of these techniques are lower than DPC, they are still high
compared to the simple zero-forcing precoder. Therefore, VP with sphere encoding or lattice reduction can only be used in systems with small number of users (e.g., < 10 users).

It is desired that large number of users be supported in practical systems. Most precoders either do not scale well in terms of complexity or show poor performance for large number of users. For large multiuser MIMO systems with tens to hundreds of users, sphere encoder or lattice reduction based VP is prohibitive. In the third part of this thesis (Chapter 7), we present a norm descent search (NDS) algorithm. We show that VP with the proposed NDS achieves good performance at low complexity in large multiuser MISO systems.

1.6 Challenges in Realizing Large-MIMO Systems

We note that current wireless standards including IEEE 802.11n (WiFi), IEEE 802.16e (WiMax) and 3GPP LTE-A have adopted MIMO techniques to achieve increased capacity/spectral efficiency and reliability. These standards harness only a limited potential of the capacity benefits of MIMO, since they currently use only a small number of transmit antennas (e.g., 2 to 4 antennas) and achieve spectral efficiencies of only about 10 bps/Hz or less. However, significant benefits can be realized if large number of antennas are used; e.g., large-MIMO systems with tens of antennas in communication terminals can enable multi-giga bit rate transmissions at high spectral efficiencies of the order of tens to hundreds of bps/Hz. Key challenges in realizing such large-MIMO systems include:

- availability of large number of independent spatial dimensions
- placement of many antennas and RF/IF chains in communication terminals
- low-complexity signal detection and channel estimation.
1.6.1 Availability of Large Number of Independent Spatial Dimensions

The number of independent spatial dimensions is limited by the amount/richness of scattering around the wireless communication terminals. Even in the presence of rich scattering, pin-hole effects can arise, where all the paths from the transmit to receive antennas go through a common pin-hole [8]. These can result in reduced number of independent spatial dimensions (i.e., low rank channel matrices). Rich scattering is however common in indoor and outdoor urban settings. In addition, spacing between the antennas in the communication terminal is also crucial in determining the number of independent spatial dimensions. In the case of large-MIMO systems, providing adequate spacing between antennas necessitates that the communication terminals be medium-/large-sized to accommodate many antennas. This point is elaborated next.

1.6.2 Placement of Large Number of Antennas

Increasing the spacing between antennas in a communication terminal is important to reduce correlation between the channel gains, since it is known that receive correlation reduces the MIMO capacity. A spacing of more than $\lambda/2$, where $\lambda$ is the carrier wavelength, is considered sufficient for achieving almost no correlation between antennas. While achieving this separation may be readily feasible at the base stations, achieving it in small-sized user terminals will be an issue. For example, at 2.5 GHz carrier frequency, half wavelength is 6 cm; therefore, at most 2 antennas can only be placed in hand held terminals like cell phones.

However, there are several applications where communication terminals can be medium-/large-sized. For example, in applications like fixed wireless IPTV/HDTV distribution, high data rate backhaul wireless links between wireless mesh nodes or base station controllers (BSC), the aperture of the communication terminals (e.g., set top boxes, laptops, TVs) is much larger, and this allows large number of antennas to be
easily placed in them. Reduced carrier wavelengths can further alleviate the antenna placement problem. For example, operation in 5 GHz ($\lambda/2$ is 3 cm) and 60 GHz ($\lambda/2$ is 2.5 mm) bands will allow more number of antennas to be placed in communication terminals.

**Some Recent Large-MIMO Channel Sounding Setups**

We highlight some recent trends in high spectral efficiency MIMO systems/channel sounding measurements with large number of antennas.

- NTT DoCoMo has reported a $12 \times 12$ V-BLAST MIMO system at 5 Gbps data rate and 50 bps/Hz spectral efficiency in 4.6 GHz band at a mobile speed of 10 Km/hr [46].

- Evolution of WiFi standards (evolution from IEEE 802.11n to IEEE 802.11ac to achieve multi-gigabit rate transmissions in 5 GHz band) now considers $16 \times 16$ MIMO operation; e.g., see $16 \times 16$ MIMO indoor channel sounding measurements at 5.17 GHz reported in [47] for consideration in WiFi standards (see Fig. 1.3).

- $64 \times 64$ MIMO channel sounding measurements at 5 GHz in indoor environments have been reported in [48] (see Fig. 1.4).

We note that, while the RF/antenna technologies/measurements for large-MIMO systems are getting matured, there is lack of current focus on development of low-complexity algorithms for detection and channel estimation for large-MIMO systems (e.g., MIMO systems with 16 or more antennas) to reap their high spectral efficiency benefits. A vast body of MIMO detection literature is heavily focused on only up to $4 \times 4$ (in some cases $8 \times 8$) MIMO. This observation brings to focus the issue of low-complexity large-MIMO detection, which is discussed next.
1.6.3 Low-Complexity Detection/Channel Estimation

A key component of a MIMO system is the MIMO detector at the receiver, whose job is to recover the symbols that are transmitted simultaneously from multiple transmitting antennas. In practical applications, the MIMO detector is often the bottleneck for the overall performance and complexity. Complexities involved in optimum detectors based on maximum likelihood (ML) or maximum a posteriori (MAP) criterion are exponential in number of transmit antennas [4],[57]. Such complexities are prohibitively high for large $N_t$.

MIMO detectors including sphere decoder and several of its variants [49]-[55] achieve ML/near-ML performance at the cost of high complexity. Other well known suboptimum detectors including ZF (zero forcing), MMSE (minimum mean square error), and ZF-SIC (ZF with successive interference cancellation) detectors [4],[57] are attractive from a complexity view point, but achieve relatively poor performance. For example,
the ZF-SIC detector (i.e., the well known V-BLAST detector with ordering [18],[58]) does not achieve the full diversity in the system; diversity achieved by ZF-SIC detector in V-BLAST is only $N_r - N_t + 1$ [1]. Also, the per-symbol complexity of ZF-SIC detector is cubic in number of transmit antennas, which is still high for large $N_t$. Therefore, there is a need for large-MIMO detectors which can perform close to the performance of the optimum detector, but at practically affordable complexities.

### Recent Developments in Large-MIMO Detection

In this context, we note that certain algorithms rooted in neural networks/machine learning/artificial intelligence have been shown to achieve near-optimal performance in multiuser detection of CDMA signals, when the number of users in the system gets large. Some of these algorithms include local neighborhood likelihood ascent search (LAS) algorithm [59], probabilistic data association (PDA) [60]-[62], and belief propagation (BP) [67]-[69]. *Motivated by the suitability of such algorithms for multiuser detection in large CDMA systems, in this thesis, we successfully develop two of these algorithms, namely, LAS and PDA algorithms, in the context of large-MIMO detection, and establish that*
near-optimal detection can be achieved in large-MIMO systems at practically affordable complexities. For e.g., the per-symbol complexity of the LAS algorithm is only quadratic in number of transmit antennas for $N_t = N_r$ V-BLAST MIMO, which makes the algorithm amenable for practical implementations. Specifically, following our reporting of the LAS algorithm for large-MIMO detection in [70],[71], an FPGA/ASIC implementation of the LAS algorithm for $32 \times 32$ V-BLAST MIMO with 4-/16-/64-QAM has been reported in [72]. We expect that investigation and reporting of more of such large-MIMO detection algorithms and implementations will follow in the coming years. For e.g., a reactive tabu search (RTS) based algorithm and two BP based algorithms for large-MIMO detection (in $32 \times 32$ and $64 \times 64$ MIMO) have been reported recently in [73],[74] and [75],[76]. Another large-MIMO detection algorithm (e.g., in $50 \times 50$ MIMO) based on Gibbs sampling has been reported recently in [77].

We further note that CSIR is needed for MIMO detection. So, accurate channel estimation at the receiver is crucial. For large-MIMO systems, the $N_tN_r$ number of channel coefficients to be estimated is large. We propose a pilot-assisted training based iterative detection/channel estimation scheme for this purpose [78].

With the feasibility of such low-complexity algorithms, detection/channel estimation need not be a bottleneck to implement large-MIMO systems. This message is one of the key contributions made in this thesis.

1.7 Problems Addressed and Contributions Made

The work reported in this thesis is comprised of the following three major parts:

1. Low-complexity detection, based on a local neighborhood search and probabilistic data association, on large-MIMO links with CSIR only, and the associated channel estimation. (Chapters 2, 3, 4).
2. Low-complexity precoding using X-Codes/X-Precoders and Y-Codes/Y-Precoders on large-MIMO links with CSIT and CSIR. (Chapters 5, 6).

3. Low-complexity precoding for large multiuser MISO downlink systems with CSIT, based on vector perturbation with a reduced search space. (Chapter 7).

1.7.1 Low-Complexity Large-MIMO Detection/Channel Estimation

In this part of the thesis, we present two low-complexity large-MIMO detection algorithms and their uncoded/coded bit error performances in i.i.d. and spatially correlated MIMO channels. The first algorithm is a multistage LAS \((M\text{-LAS})\) algorithm based on a local neighborhood search with ML cost as the search metric \([70],[71],[79]-[82]\). The second algorithm is the PDA algorithm based on MAP criterion \([83]\). The performance of these two algorithms are evaluated for the two popular MIMO architectures, namely, \(i\) V-BLAST MIMO, and \(ii\) non-orthogonal STBC MIMO. While V-BLAST MIMO exploits only spatial dimensions to achieve full rate (i.e., maximum multiplexing gain of \(\min(N_t, n_r)\)), fully diverse non-orthogonal STBC MIMO \([21]\) exploits spatial and time dimensions to achieve both full diversity \((N_t N_r)\) as well as full rate.

**Multistage Likelihood Ascent Search**

The LAS algorithm starts with an initial solution vector, which can be, for e.g., MF solution vector or ZF solution vector or MMSE solution vector. A neighborhood around the initial vector is defined; e.g., set of all vectors which differ from the initial solution vector in one coordinate is an example of a neighborhood. For each of the vectors in the neighborhood, the algorithm computes the ML cost function. The best vector among the neighboring vectors (in terms of least ML cost among them) which also happens to have a lesser ML cost than that of the initial vector is chosen, and declared as the new solution vector. This new solution vector is passed on as the initial vector for the next
iteration, where the best vector among the neighboring vectors of the current initial vector is chosen as the new solution vector for the next iteration, and so on until a local minima is reached. The algorithm ends once a local minima is encountered, and the local minima is declared as the final solution vector.

**LAS Algorithm Complexity:**

A key advantage of the LAS algorithm is its simplicity in its search operation. Much of the algorithm complexity arises from the initial vector computation (which involves matrix inversion operation in ZF and MMSE solutions) and the computation of $H^T H$, which requires $O(N_t N_r)$ complexity per symbol. The average per-symbol complexity in the search part alone is found to be $O(N_t)$ through simulations. So the overall per-symbol complexity of the LAS algorithm is $O(N_t N_r)$. This low order of complexity is well suited for scaling to large number of dimensions.

**LAS Algorithm Performance:**

A even more interesting aspect of the LAS algorithm is that its bit error rate (BER) performance improves with increasing values of $N_t = N_r$ in V-BLAST MIMO; a behavior we refer to as the ‘large-system behavior’ of the algorithm. Increasingly closer to ML performance is achieved for increasing number of transmit antennas.

**Applicability to Large Non-Orthogonal STBC MIMO Systems:**

We note that large number of dimensions are required for achieving near-optimal performance with LAS. Since STBCs code across both space and time, it is possible to achieve large number of dimensions in STBC MIMO systems with lesser number of antennas as compared to V-BLAST MIMO systems. For example, hundreds of dimensions can be created with tens of dimensions each in space and time using non-orthogonal STBCs which achieve both full rate (same as that achieved in V-BLAST) as well as full transmit diversity [21]. For example, a $16 \times 16$ non-orthogonal STBC matrix in [21] is constructed using 256 complex data symbols resulting in 512 real dimensions; with 64-QAM and rate-3/4 turbo code, this STBC achieves a spectral efficiency of 72
bps/Hz. In Chapter 2, we establish through extensive simulations that the LAS algorithm is very effective, both in terms of complexity as well as achieving near-ML/near capacity performance, in decoding $16 \times 16$ and $32 \times 32$ non-orthogonal STBCs, even in the presence of spatial correlation and with estimated channel matrix.

**Multistage LAS:**

In an attempt to improve the performance of the basic LAS algorithm, we propose a more general version of LAS algorithm, termed as multistage LAS algorithm. This algorithm executes an escape mechanism when it encounters a local minima, by changing the neighborhood definition: it considers vectors which differ in two or more coordinates (as opposed to only one coordinate in the basic neighborhood definition in LAS) as neighbors. On escaping from a local minima, the algorithm reverts back to the basic neighborhood definition till the next local minima is encountered and stops when no escape from a local minima is possible.

Our contributions in this part can be summarized as follows [70],[71],[78]-[82]:

- We develop the basic LAS algorithm with 1-symbol neighborhood for use in V-BLAST and non-orthogonal STBC MIMO systems. We show that LAS detection of $64 \times 64$ and $128 \times 128$ V-BLAST MIMO signals (i.e., $N_t = N_r = 64, 128$) achieves close to SISO AWGN performance\(^1\). In a $128 \times 128$ V-BLAST system with 4-QAM, LAS algorithm achieves an uncoded BER of $10^{-3}$ at an SNR of just about 1 dB away from SISO AWGN performance. In terms of coded BER, with a rate-3/4 turbo code at a spectral efficiency of 192 bps/Hz, the algorithm performs close to within about 4.5 dB from theoretical MIMO capacity.

- We generalize the basic 1-symbol neighborhood LAS algorithm by employing a low-complexity multistage multi-symbol neighborhood based strategy; we refer

---

\(^1\)Since simulation of brute-force ML or sphere decoder is prohibitively complex for such large dimensions, we use the SISO AWGN performance as a lower bound on the true ML performance for comparison purposes.
to this as multistage LAS (M-LAS) algorithm. We show that the M-LAS algorithm outperforms the basic LAS algorithm with some increase in complexity.

- We propose a method to generate soft outputs from the M-LAS output vector. The proposed soft outputs generation for the individual bits results in about 1 to 1.5 dB improvement in coded BER compared to hard decision M-LAS outputs.

- Assuming i.i.d. fading and perfect CSIR, our simulation results show that the proposed M-LAS algorithm is able to decode large non-orthogonal STBCs (e.g., $16 \times 16$ and $32 \times 32$ STBCs) and achieve near SISO AWGN uncoded BER performance as well as near-capacity (within 4 dB from theoretical MIMO capacity) coded BER performance.

- Using the proposed detector, we decode and report the simulated BER performance of ‘perfect codes’ [22],[84]-[87] of large dimensions.

- Presenting a BER performance and complexity comparison of the proposed non-orthogonal STBC/LAS detection approach with other large-MIMO/detector approaches (e.g., stacked Alamouti codes/QOSTBCs and associated interference canceling receivers reported in [88]), we show that the proposed approach outperforms the other considered approaches, both in terms of performance as well as complexity.

- We present simulation results that quantify the loss in BER performance due to spatial correlation in large-MIMO systems, by considering a more realistic spatially correlated MIMO fading channel model proposed by Gesbert et al in [8]. We show that this loss in performance can be alleviated by providing more receive dimensions (i.e., more receive antennas than transmit antennas).

- We present a training-based iterative detection/channel estimation scheme for
large non-orthogonal STBC MIMO systems. We report BER and nearness-to-capacity results when the channel matrix is estimated using the proposed iterative scheme and compare these results with those obtained using perfect CSIR assumption.

- We present an asymptotic \((N_t, N_r \to \infty, \text{keeping } N_t = N_r)\) performance analysis of the basic LAS algorithm detection in V-BLAST MIMO with 4-QAM in i.i.d. Rayleigh fading with a motivation to get some insights that can explain the good performance of the LAS algorithm in large dimensions.

**Probabilistic Data Association**

PDA algorithm was originally developed for target tracking. It is widely used in digital communications [60]-[66]. Particularly, PDA algorithm is a reduced complexity alternative to the a posteriori probability (APP) decoder/detector/equalizer. Near-optimal performance has been demonstrated for PDA-based multiuser detection in CDMA systems with large number of users [60]-[62]. PDA has been used in the detection of V-BLAST signals with small number of antennas [64]-[66]. Here, we develop the PDA algorithm for use in large V-BLAST as well as non-orthogonal STBC MIMO detection, and present its uncoded and coded BER performance [83]. The complexity of PDA is more than that of LAS. PDA is shown to perform better than LAS in higher-order QAM (e.g., 16-QAM) at low SNRs.

**1.7.2 Low-Complexity Large-MIMO Precoding Using X-, Y-Codes**

In this part of the thesis, we consider MIMO systems where CSI is fully available both at the transmitter and the receiver. It is known that precoding techniques can provide large performance improvements in such scenarios. A popular precoding approach is based on singular value decomposition (SVD) [89],[90] of the channel so that the MIMO channel can be seen as parallel channels.
In slow fading scenarios, channels are subject to block fading. Without rate and power adaptation, outages cannot be avoided. In such scenarios, a popular measure of reliability is the diversity order achieved by a given transmit-receive scheme. We consider SVD precoding for MIMO systems, which transforms the MIMO channels into parallel subchannels. At the receiver, ML decoding (MLD) can be employed separately for each subchannel. To improve the low diversity order of the SVD precoded system, we propose some simple linear codes prior to SVD precoding. These codes are named X- and Y-Codes due to the structure of the encoder matrix, which enables us to flexibly pair subchannels with different diversity orders. Specifically, the subchannels with low diversity orders can be paired together with those having high diversity orders, so that the overall diversity order is improved. The main contributions in this part are [93],[94]:

1. **X-Codes:** A set of 2-dimensional (2-D) real orthogonal matrices is used to jointly code over pairs of subchannels, without increasing the transmit power. Since the matrices are effectively parameterized with a single angle, the design of X-Codes primarily involves choosing the optimal angle for each pair of subchannels. The angles are chosen *a priori* and do not change with each channel realization. This is why we use the term ‘Code’ instead of ‘Precoder’. Further optimization of angles are based upon minimizing the average error performance. At the receiver, we show that the MLD can be easily accomplished using \( N \), low complexity 2-D real sphere decoders (SD) [49]. It is shown that X-Codes have better error performance than that of other precoders, yet it becomes worse when the pair of subchannels is poorly conditioned. This motivates us to propose Y-Codes.

2. **Y-Codes:** Instead of using rotation for pairing subchannels, we use a linear code generator matrix which is upper left triangular. Y-Codes are parameterized with 2 parameters corresponding to power allocated to the two subchannels. These parameters are computed so as to minimize the average error probability. The
MLD complexity is the same as that of the scalar channels in linear precoders [25],[26] and is less than that of the X-Codes, while the performance of Y-Codes is better than that of X-Codes for ill-conditioned channel pairs.

3. **X-, Y-Precoders:** The X- and Y-Precoders employ the same pairing structure as that in X-, Y-Codes. However, the code generator matrix for each pair of subchannels is chosen for each channel realization. We observed that the error performance of X- and Y-Precoders is better than that of X- and Y-Codes.

**Precoding Using X-Codes to Increase MIMO Capacity with Discrete Alphabets**

It is known that the capacity of the Gaussian MIMO channel with CSIT can be achieved with Gaussian inputs. However, in practice, the input alphabet is *not Gaussian* and is generally chosen from a finite signal set. Therefore, precoders should be designed to achieve the capacity of the Gaussian MIMO channel with discrete input alphabets. The optimal precoder with discrete inputs is given by a fixed point equation, which requires a high complexity numerical evaluation [95]. Since the optimal precoder jointly codes all the $N_t$ inputs, joint decoding is also required at the receiver. Thus, the decoding complexity can be very high, specially for large $N_t$. Motivated by this issue of high complexity of the optimal precoder with discrete inputs, we propose a precoder based on X-Codes, which is shown to achieve mutual information close to the discrete input MIMO capacity, at low complexity [96],[97].

The structure of X-Codes enables us to express the total mutual information as a sum of the mutual information of all the pairs of subchannels. The problem of finding the optimal precoder with the above structure, which maximizes the total mutual information, is solved by i) optimizing the rotation angle and the power allocation within each pair, and ii) finding the optimal pairing and power allocation among the pairs. It is shown that the mutual information achieved with the proposed pairing scheme is very close to that achieved with the optimal precoder by Cruz *et al.* [95], and is
significantly better than Mercury/waterfilling strategy by Lozano et al. [98]. Our approach greatly simplifies both the precoder optimization and the detection complexity, making it suitable for practical applications.

1.7.3 Low-Complexity Large Multiuser MISO Precoding

In this part of the work, we consider the problem of precoding in large multiuser MISO systems with large number of transmit antennas \( N_t \) at the base station and large number of downlink users \( N_u \), where each user has one receive antenna. Such large MISO systems are of interest because of the high capacities (sum-rates) of the order of tens to hundreds of bits/channel use possible in such systems. We propose a vector perturbation based low-complexity precoder, termed as norm descent search (NDS) precoder [99], which has a complexity of just \( O(N_u N_t) \) per information symbol. This low complexity attribute of the precoder is achieved by searching for the perturbation vector over a reduced search space. Interestingly, in terms of BER performance, the proposed precoder achieves increasingly better BER for increasing \( N_t, N_u \), making it suited for large MISO systems both in terms of complexity as well as performance.

1.8 Organization of the Thesis

The rest of the thesis is organized as follows. In Chapter 2, we present the proposed M-LAS algorithm and its uncoded and turbo coded BER performance in large V-BLAST MIMO and non-orthogonal STBC MIMO systems. Performance of the M-LAS algorithm with estimated CSIR using an iterative detection/channel estimation scheme is also presented in this chapter. In Chapter 3, we present the asymptotic performance analysis of the LAS algorithm. In Chapter 4, we present the PDA algorithm, its BER performance and complexity in large-MIMO detection. In Chapter 5, we present the
proposed large-MIMO precoding schemes using X-Codes/Precoders and Y-Codes/Precoders and their performance. In Chapter 6, the precoder based on X-Codes to increase the mutual information in Gaussian MIMO channels with discrete input alphabets is presented. In Chapter 7, the proposed vector perturbation based large multiuser MISO precoding scheme that uses a low-complexity norm descent search is presented. Finally, conclusions are presented in Chapter 8.
Chapter 2

Large-MIMO Detection Using Likelihood Ascent Search

In this Chapter, we present a low-complexity detection algorithm, termed as likelihood ascent search (LAS) algorithm, suited for large-MIMO detection, and evaluate its uncoded and coded BER performance [70],[71],[78]-[82]. The LAS algorithm is a local neighborhood search algorithm. The neighborhood of a initial solution vector in a given iteration of the algorithm is defined as a collection of those vectors which differ in certain number of coordinates compared to the initial solution vector. Better solution vectors compared to the initial solution vector (in terms of maximum-likelihood cost) in this neighborhood are searched for. The best among the neighbors is fed as the initial solution vector for the next iteration. The iterations are continued till a local minima is reached, upon which the local minima is declared as the solution vector, or an escape strategy is adopted to leave the local minima and continue the search for better local minima. The LAS algorithm is shown to exhibit large-system effect, where the BER performance improves with increasing number of antennas. We also show that the algorithm achieves near-capacity performance within about 4 dB from the theoretical limit. We investigate the effect of spatial correlation on the BER performance of the LAS algorithm. An iterative detection/channel estimation scheme and its BER
performance are also presented.

This chapter is organized as follows. In Section 2.1, we present the system models for V-BLAST MIMO and non-orthogonal STBC MIMO. In Section 2.2, the proposed multistage LAS algorithm is presented. The BER performance of LAS detection in large V-BLAST MIMO and non-orthogonal STBC MIMO are presented in Sections 2.3 and 2.4, respectively. The proposed iterative detection/channel estimation scheme and its performance are presented in Section 2.5.

2.1 System Model

In this section, we present system models corresponding to V-BLAST MIMO and non-orthogonal STBC MIMO, and a unified system model covering both models so that the LAS detection algorithm can be presented under the unified framework.

2.1.1 V-BLAST MIMO

Consider a V-BLAST system with \( N_t \) transmit antennas and \( N_r \) receive antennas, \( N_t \leq N_r \), where \( N_t \) symbols are transmitted from \( N_t \) transmit antennas simultaneously. Let \( x_c \in \mathbb{C}^{N_t} \) be the symbol vector transmitted. Each element of \( x_c \) is an \( \mathcal{M} \)-PAM or \( \mathcal{M} \)-QAM symbol. Let \( H_c \in \mathbb{C}^{N_r \times N_t} \) be the channel gain matrix, such that the \((p, q)\)th entry \( h_{p,q} \) is the complex channel gain from the \( q \)th transmit antenna to the \( p \)th receive antenna. Assuming rich scattering, we model the entries of \( H_c \) as i.i.d \( \mathcal{CN}(0, 1) \). Let \( y_c \in \mathbb{C}^{N_r} \) and \( n_c \in \mathbb{C}^{N_r} \) denote the received signal vector and the noise vector, respectively, at the receiver, where the entries of \( n_c \) are modeled as i.i.d \( \mathcal{CN}(0, \sigma^2) \). The received signal vector can then be written as

\[
y_c = H_c x_c + n_c.
\] (2.1)
Let $y_c, H_c, x_c$, and $n_c$ be decomposed into real and imaginary parts as follows:

$$
y_c = y_I + jy_Q, \quad x_c = x_I + jx_Q, \quad n_c = n_I + jn_Q, \quad H_c = H_I + jH_Q.
$$

Further, we define $H_r \in \mathbb{R}^{2N_r \times 2N_t}, y_r \in \mathbb{R}^{2N_r}, x_r \in \mathbb{R}^{2N_t}$, and $n_r \in \mathbb{R}^{2N_r}$ as

$$
H_r \triangleq \begin{pmatrix} H_I & -H_Q \\ H_Q & H_I \end{pmatrix}, \quad y_r = \begin{bmatrix} y_I^T \\ y_Q^T \end{bmatrix}, \quad x_r = \begin{bmatrix} x_I^T \\ x_Q^T \end{bmatrix}, \quad n_r = \begin{bmatrix} n_I^T \\ n_Q^T \end{bmatrix}.
$$

Now, (2.1) can be written as

$$
y_r = H_r x_r + n_r.
$$

We will work with the real-valued system in (2.3). For notational simplicity, we drop subscripts $r$ in (2.3) and write

$$
y = Hx + n,
$$

where $H = H_r \in \mathbb{R}^{2N_r \times 2N_t}, y = y_r \in \mathbb{R}^{2N_r}, x = x_r \in \mathbb{R}^{2N_t}$ and $n = n_r \in \mathbb{R}^{2N_r}$.

With the above real-valued system model, the real-part of the original complex data symbols will be mapped to $[x_1, \ldots, x_{N_t}]$ and the imaginary-part of these symbols will be mapped to $[x_{N_t+1}, \ldots, x_{2N_t}]$.

### 2.1.2 Non-Orthogonal STBC MIMO

Consider a STBC MIMO system with multiple transmit and multiple receive antennas. An $(n, p, k)$ STBC is represented by a matrix $X_c \in \mathbb{C}^{n \times p}$, where $n$ and $p$ denote the number of transmit antennas and number of time slots, respectively, and $k$ denotes the number of complex data symbols sent in one STBC matrix [2]. The $(i, j)$th entry in $X_c$ represents the complex number transmitted from the $i$th transmit antenna in the $j$th time slot. The rate of an STBC, $r$, is given by $r \triangleq \frac{k}{p}$. Let $N_r$ and $N_t = n$ denote the
number of receive and transmit antennas, respectively. Let $H_c \in \mathbb{C}^{N_r \times N_t}$ denote the channel gain matrix, where the $(i, j)$th entry in $H_c$ is the complex channel gain from the $j$th transmit antenna to the $i$th receive antenna. We assume that the channel gains remain constant over one STBC matrix duration. Assuming rich scattering, we model the entries of $H_c$ as i.i.d $\mathcal{CN}(0, 1)$. The received space-time signal matrix, $Y_c \in \mathbb{C}^{N_r \times p}$, can be written as

$$Y_c = H_c X_c + N_c,$$  \hspace{1cm} (2.5)

where $N_c \in \mathbb{C}^{N_r \times p}$ is the noise matrix at the receiver and its entries are modeled as i.i.d $\mathcal{CN}(0, \sigma^2)$. The $(i, j)$th entry in $Y_c$ is the received signal at the $i$th receive antenna in the $j$th time slot. In a linear dispersion (LD) STBC, $X_c$ can be decomposed into a linear combination of weight matrices corresponding to each data symbol and its conjugate as [2]

$$X_c = \sum_{i=1}^{k} x_c^{(i)} A_c^{(i)} + (x_c^{(i)})^* E_c^{(i)},$$  \hspace{1cm} (2.6)

where $x_c^{(i)}$ is the $i$th complex data symbol, and $A_c^{(i)}, E_c^{(i)} \in \mathbb{C}^{N_t \times p}$ are its corresponding weight matrices. The detection algorithm we present in this Chapter can decode general LD STBCs of the form in (2.6). For the purpose of simplicity in exposition, here we consider a subclass of LD STBCs, where $X_c$ can be written in the form

$$X_c = \sum_{i=1}^{k} x_c^{(i)} A_c^{(i)}.$$  \hspace{1cm} (2.7)

From (2.5) and (2.7), applying the $\text{vec} \, (.)$ operation we have

$$\text{vec} (Y_c) = \sum_{i=1}^{k} x_c^{(i)} \text{vec} (H_c A_c^{(i)}) + \text{vec} (N_c).$$  \hspace{1cm} (2.8)
If $U,V,W,D$ are matrices such that $D = UWV$, then it is true that $vec(D) = (V^T \otimes U) vec(W)$, where $\otimes$ denotes tensor product of matrices [108]. Using this, we can write (2.8) as

$$vec(Y_c) = \sum_{i=1}^{k} x_c^{(i)} (I_p \otimes H_c) vec(A_c^{(i)}) + vec(N_c).$$  (2.9)

Further, define $y_c \triangleq vec(Y_c)$, $\hat{H}_c \triangleq (I \otimes H_c)$, $a_c^{(i)} \triangleq vec(A_c^{(i)})$, and $n_c \triangleq vec(N_c)$. From these definitions, it is clear that $y_c \in \mathbb{C}^{N_p}$, $\hat{H}_c \in \mathbb{C}^{N_p \times N_p}$, $a_c^{(i)} \in \mathbb{C}^{N_p}$, and $n_c \in \mathbb{C}^{N_p}$.

Let us also define a matrix $\tilde{H}_c \in \mathbb{C}^{N_p \times k}$, whose $i$th column is $\hat{H}_c a_c^{(i)}$, $i = 1, \ldots, k$. Let $x_c \in \mathbb{C}^k$, whose $i$th entry is the data symbol $x_c^{(i)}$. With these definitions, we can write (2.9) as

$$y_c = \sum_{i=1}^{k} x_c^{(i)} (\hat{H}_c a_c^{(i)}) + n_c = \tilde{H}_c x_c + n_c.$$  (2.10)

Let $y_c, \tilde{H}_c, x_c$, and $n_c$ be decomposed into real and imaginary parts as

$$y_c = y_I + jy_Q, \quad x_c = x_I + jx_Q, \quad n_c = n_I + jn_Q, \quad \tilde{H}_c = \tilde{H}_I + j\tilde{H}_Q.$$  (2.11)

Further, we define $x_r \in \mathbb{R}^{2k}, y_r \in \mathbb{R}^{2N_p}, H_r \in \mathbb{R}^{2N_p \times 2k}$, and $n_r \in \mathbb{R}^{2N_p}$ as

$$H_r = \begin{pmatrix} \tilde{H}_I & -\tilde{H}_Q \\ \tilde{H}_Q & \tilde{H}_I \end{pmatrix}, \quad n_r = [n_I^T \ n_Q^T]^T, \quad x_r = [x_I^T \ x_Q^T]^T, \quad y_r = [y_I^T \ y_Q^T]^T.$$  (2.12)

Now, (2.10) can be written as

$$y_r = H_r x_r + n_r.$$  (2.13)
Chapter 2. Large-MIMO Detection Using Likelihood Ascent Search

We will work with the real-valued system in (2.13). For notational simplicity, we drop subscripts $r$ in (2.13) and write

$$
y = Hx + n, \quad (2.14)$$

where $H = H_r \in \mathbb{R}^{2N_r \times 2k}$, $y = y_r \in \mathbb{R}^{2N_r \times p}$, $x = x_r \in \mathbb{R}^{2k}$, and $n = n_r \in \mathbb{R}^{2N_r \times p}$.

### 2.1.3 High-rate Non-orthogonal STBCs from CDA

We focus on the detection of square (i.e., $n = p = N_t$), full-rate (i.e., $k = pn = N_t^2$), circulant (where the weight matrices $A_c^{(i)}$'s are permutation type), non-orthogonal STBCs from CDA, whose construction for arbitrary number of transmit antennas $n$ is given by the matrix in (2.15) given by [21]:

$$X_c = \begin{bmatrix}
\sum_{i=0}^{n-1} x_{0,i} t^i & \delta \sum_{i=0}^{n-1} x_{0,1,i} \omega_n^i t^i & \delta \sum_{i=0}^{n-1} x_{0,2,i} \omega_n^{2i} t^i & \cdots & \delta \sum_{i=0}^{n-1} x_{0,n-1,i} \omega_n^{(n-1)i} t^i \\
\sum_{i=0}^{n-1} x_{1,i} t^i & \delta \sum_{i=0}^{n-1} x_{1,1,i} \omega_n^i t^i & \delta \sum_{i=0}^{n-1} x_{1,2,i} \omega_n^{2i} t^i & \cdots & \delta \sum_{i=0}^{n-1} x_{1,n-1,i} \omega_n^{(n-1)i} t^i \\
\sum_{i=0}^{n-1} x_{2,i} t^i & \sum_{i=0}^{n-1} x_{2,1,i} \omega_n^i t^i & \sum_{i=0}^{n-1} x_{2,2,i} \omega_n^{2i} t^i & \cdots & \sum_{i=0}^{n-1} x_{2,n-1,i} \omega_n^{(n-1)i} t^i \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\sum_{i=0}^{n-1} x_{n-1,i} t^i & \sum_{i=0}^{n-1} x_{n-1,1,i} \omega_n^i t^i & \sum_{i=0}^{n-1} x_{n-1,2,i} \omega_n^{2i} t^i & \cdots & \sum_{i=0}^{n-1} x_{n-1,n-1,i} \omega_n^{(n-1)i} t^i \\
\end{bmatrix}, \quad (2.15)
$$

In (2.15), $\omega_n = e^{\frac{2\pi i}{n}}$, and $x_{u,v}$, $0 \leq u, v \leq n - 1$ are the data symbols from a QAM alphabet. When $\delta = e^{\sqrt{3} j}$ and $t = e^j$, the STBC in (2.15) achieves full transmit diversity (under ML decoding) as well as information-losslessness [21]. When $\delta = t = 1$, the code ceases to be of full-diversity (FD), but continues to be information-lossless (ILL) [109],[110]. High spectral efficiencies with large $n$ can be achieved using this code construction. For example, with $n = 32$ transmit antennas, the $32 \times 32$ STBC from (2.15) with 16-QAM and rate-3/4 turbo code achieves a spectral efficiency of 96 bps/Hz. This...
high spectral efficiency is achieved along with the full-diversity of order \( nN_r \). However, since these STBCs are non-orthogonal, ML detection gets increasingly impractical for large \( n \). Consequently, a key challenge in realizing the benefits of these large STBCs in practice is that of achieving near-ML performance for large \( n \) at low detection complexities. The multistage likelihood ascent search (M-LAS) detector proposed in Sec. 2.2 essentially addresses this challenging issue.

### 2.1.4 Unified System Model

A unified linear vector channel model for both V-BLAST MIMO and non-orthogonal STBC MIMO is given by

\[
y = Hx + n, \tag{2.16}
\]

where \( H \in \mathbb{R}^{2N_r,p \times 2k} \) is the equivalent matrix of channel gains, \( y \in \mathbb{R}^{2N_r,p} \) is the equivalent received signal vector, \( x \in \mathbb{R}^{2k} \) is the equivalent transmitted symbol vector, and \( n \in \mathbb{R}^{2N_r,p} \) is the equivalent additive white Gaussian noise vector. We can get the V-BLAST MIMO system model by simply substituting \( p = 1 \) and \( k = N_t \) in (2.16). Similarly, we can get the non-orthogonal STBC MIMO system model by substituting \( p = N_t \) and \( k = N_t^2 \).

The transmitted symbols are assumed to be \( \mathcal{M} \)-QAM. For simplicity, we consider square-QAM. However, the algorithm is valid for rectangular QAM as well. The \( i \)th element of \( x \), \( x_i \), is therefore a \( \sqrt{\mathcal{M}} \)-PAM symbol and takes discrete values from the set \( A_i = \{ A_m, m = 1, \ldots, \sqrt{\mathcal{M}} \} \), where \( A_m = (2m - 1 - \sqrt{\mathcal{M}}) \). In both V-BLAST MIMO and non-orthogonal STBC MIMO, the average signal power received at each receive antenna is \( N_t E_s \), where \( E_s \) is the average energy of each transmitted symbol. Therefore, the average SNR at each receive antenna is given by \( \gamma = \frac{N_t E_s}{\sigma^2} \) [2]. Now, define a \( 2k \)-dimensional signal space \( S \) to be the Cartesian product of \( A_1 \) to \( A_{2k} \).
2.1.5 Maximum-Likelihood Detector

The maximum-likelihood (ML) detector is the optimal detector w.r.t minimizing the word error probability. One word corresponds to one $N_t$ dimensional transmitted symbol vector in case of V-BLAST. In the case of non-orthogonal STBC MIMO, one word corresponds to one $N_t \times N_t$ STBC codeword matrix. The ML decision rule is given by

$$d_{ML} = \arg \min_{d \in S} \left\| y - Hd \right\|^2 = \arg \min_{d \in S} d^T H^T H d - 2y^T H d,$$

(2.17)

whose complexity is exponential in $k$ [57]. So, detection algorithms which are low in complexity but close to ML in terms of performance are of interest in large-MIMO systems, where $k$ is large.

ML Performance in Large Dimensions: Since we are interested in detectors with performance close to the ML performance, it would be necessary to know the true ML performance for comparison purposes. However, exact analytical expressions do not exist for the word or bit error probability of V-BLAST and non-orthogonal STBC MIMO with ML detection. It is also difficult to know the error probabilities from Monte-Carlo simulation of ML detection (except for systems with small dimensions) due to the exponential complexity in number of dimensions. Even sphere decoder, which is known to achieve ML performance, and its low-complexity variants do not scale well for large dimensions. Further, ML detector is optimal w.r.t. the word error probability. In many applications, bit and symbol error rates are more relevant performance measures. The detector which achieves minimum BER is called as the minimum BER detector [57]. Just like for ML detection, exact analytical expressions do not exist for the minimum BER detector. Because of this (i.e., due to lack of low-complexity methodologies to evaluate the exact ML or minimum BER detectors’ performance for large dimensions),
instead of comparing with the exact BER performance of ML/minimum BER detectors, we will compare the BER performance of the proposed detectors with the BER performance on single-input single-output (SISO) unfaded AWGN channel, which is a lower bound on the BER performance in MIMO fading channels.

### 2.2 Proposed Multistage LAS Detector

In this section, we present the proposed multistage LAS (M-LAS) algorithm for large-MIMO detection. The M-LAS algorithm consists of a sequence of likelihood-ascent search stages, where the likelihood increases monotonically with every search stage. Each search stage consists of several sub-stages. There can be at most $M$ sub-stages, each consisting of one or more iterations (the first sub-stage can have one or more iterations, whereas all the other sub-stages can have at most one iteration). In the first sub-stage, the algorithm updates one symbol per iteration such that the likelihood monotonically increases from one iteration to the next until a local minima is reached. Upon reaching this local minima, the algorithm initiates the second sub-stage. In the second sub-stage, a 2-symbol update is tried to further increase the likelihood. If the algorithm succeeds in increasing the likelihood by 2-symbol update, it starts the next search stage. If the algorithm does not succeed in the second sub-stage, it goes to the third sub-stage where a 3-symbol update is tried to further increase the likelihood. Essentially, in the $K$th sub-stage, a $K$-symbol update is tried to further increase the likelihood. This goes on until a) either the algorithm succeeds in the $K$th sub-stage for some $K \leq M$ (in which case a new search stage is initiated), or b) the algorithm terminates.

The M-LAS algorithm starts with an initial solution $d^{(0)}$, given by $d^{(0)} = B y$, where $B$ is the initial solution filter, which can be a matched filter (MF) or zero-forcing (ZF) filter or MMSE filter. The index $m$ in $d^{(m)}$ denotes the iteration number in a sub-stage of a given search stage. The ML cost function after the $k$th iteration in a given search
stage is

\[ C^{(k)} = d^{(k)^T} H^T H d^{(k)} - 2y^T H d^{(k)}. \]  

(2.18)

### 2.2.1 One-symbol Update

In the \((k + 1)\)th iteration, a one-symbol update tries to increase the likelihood by updating exactly one entry of the current data vector \(d^{(k)}\). Let us assume that we update the \(p\)th symbol in the \((k + 1)\)th iteration; in V-BLAST \(p\) can take value from \(1, \ldots, N_t\) for \(\mathcal{M}\)-PAM and \(1, \ldots, 2N_t\) for \(\mathcal{M}\)-QAM. The update rule can be written as

\[ d^{(k+1)} = d^{(k)} + \lambda_p^{(k)} e_p, \]  

(2.19)

where \(e_p\) denotes the unit vector with its \(p\)th entry only as one, and all other entries as zero. Also, for any iteration \(k\), \(d^{(k)}\) should belong to the space \(\mathcal{S}\), and therefore \(\lambda_p^{(k)}\) can take only certain integer values. For example, in case of 4-PAM or 16-QAM (both have the same signal set \(A_p = \{-3, -1, 1, 3\}\)), \(\lambda_p^{(k)}\) can take values only from \([-6, -4, -2, 0, 2, 4, 6]\). Using (2.18) and (2.19), and defining a matrix \(G\) as

\[ G \triangleq H^T H, \]  

(2.20)

we can write the cost difference as

\[ \Delta C_p^{k+1} \triangleq C^{(k+1)} - C^{(k)} = \lambda_p^{(k)^2} (G)_{p,p} - 2\lambda_p^{(k)} z_p^{(k)}, \]  

(2.21)

where \(h_p\) is the \(p\)th column of \(H\), \(z^{(k)} = H^T (y - H d^{(k)})\), \(z_p^{(k)}\) is the \(p\)th entry of the \(z^{(k)}\) vector, and \((G)_{p,p}\) is the \(p, p\)th entry of the \(G\) matrix. Also, let us define \(a_p\) and \(l_p^{(k)}\) as

\[ a_p = (G)_{p,p}, \quad l_p^{(k)} = |\lambda_p^{(k)}|. \]  

(2.22)
With the above variables defined, we can rewrite (2.21) as

\[ \Delta C_p^{k+1} = l_p^{(k)} a_p - 2l_p^{(k)} |z_p^{(k)}| \operatorname{sgn}(\lambda_p^{(k)}) \operatorname{sgn}(z_p^{(k)}), \]  

(2.23)

where \( \operatorname{sgn}(.) \) denotes the signum function. For the ML cost function to reduce from the \( k \)-th to the \((k+1)\)-th iteration, the cost difference should be negative. Using this fact and that \( a_p \) and \( l_p^{(k)} \) are non-negative quantities, we can conclude from (2.23) that the sign of \( \lambda_p^{(k)} \) must satisfy

\[ \operatorname{sgn}(\lambda_p^{(k)}) = \operatorname{sgn}(z_p^{(k)}). \]  

(2.24)

Using (2.24) in (2.23), the ML cost difference can be rewritten as

\[ F(l_p^{(k)}) \triangleq \Delta C_p^{k+1} = l_p^{(k)} a_p - 2l_p^{(k)} |z_p^{(k)}|. \]  

(2.25)

For \( F(l_p^{(k)}) \) to be non-positive, the necessary and sufficient condition from (2.25) is that

\[ l_p^{(k)} < \frac{2|z_p^{(k)}|}{a_p}. \]  

(2.26)

However, we can find the value of \( l_p^{(k)} \) which satisfies (2.26) and at the same time gives the largest descent in the ML cost function from the \( k \)-th to the \((k+1)\)-th iteration (when symbol \( p \) is updated). Also, \( l_p^{(k)} \) is constrained to take only certain integer values, and therefore the brute-force way to get optimum \( l_p^{(k)} \) is to evaluate \( F(l_p^{(k)}) \) at all possible values of \( l_p^{(k)} \). This would become computationally expensive as the constellation size \( M \) increases. However, for the case of 1-symbol update, we could obtain a closed-form expression for the optimum \( l_p^{(k)} \) that minimizes \( F(l_p^{(k)}) \), which is given by (corresponding theorem and proof are given in the Appendix A)

\[ l_p^{(k)}_{\text{opt}} = 2 \left[ \frac{|z_p^{(k)}|}{2a_p} \right], \]  

(2.27)
where \( \lfloor x \rfloor \) denotes the rounding operation, where for a real number \( x \), \( \lfloor x \rfloor \) is the integer closest to \( x \). If the \( p \)th symbol in \( d^{(k)} \), i.e., \( d_p^{(k)} \), were indeed updated, then the new value of the symbol would be given by

\[
\tilde{d}_p^{(k+1)} = d_p^{(k)} + l_p^{(k)} \text{sgn}(z_p^{(k)}).
\] (2.28)

However, \( \tilde{d}_p^{(k+1)} \) can take values only in the set \( \mathbb{A}_p \), and therefore we need to check for the possibility of \( \tilde{d}_p^{(k+1)} \) being greater than \( (M - 1) \) or less than \( -(M - 1) \). If \( \tilde{d}_p^{(k+1)} > (M - 1) \), then \( l_p^{(k)} \) is adjusted so that the new value of \( \tilde{d}_p^{(k+1)} \) with the adjusted value of \( l_p^{(k)} \) using (2.28) is \( (M - 1) \). Similarly, if \( \tilde{d}_p^{(k+1)} < -(M - 1) \), then \( l_p^{(k)} \) is adjusted so that the new value of \( \tilde{d}_p^{(k+1)} \) is \( -(M - 1) \). Let \( \tilde{l}_p^{(k)} \) be obtained from \( l_p^{(k, \text{opt})} \) after these adjustments.

It can be shown that if \( F(l_p^{(k)}) \) is non-positive, then \( F(\tilde{l}_p^{(k)}) \) is also non-positive. We compute \( F(l_p^{(k)}), \forall p = 1, \ldots, 2N_t^2 \). Now, let

\[
s = \arg \min_p F(\tilde{l}_p^{(k)}).
\] (2.29)

If \( F(\tilde{l}_s^{(k)}) < 0 \), the update for the \( (k + 1) \)th iteration is

\[
d^{(k+1)} = d^{(k)} + \tilde{l}_s^{(k)} \text{sgn}(z_s^{(k)}) e_s,
\] (2.30)

\[
z^{(k+1)} = z^{(k)} - \tilde{l}_s^{(k)} \text{sgn}(z_s^{(k)}) g_s,
\] (2.31)

where \( g_s \) is the \( s \)th column of \( G \). The update in (2.31) follows from the definition of \( z^{(k)} \) in (2.21). If \( F(\tilde{l}_s^{(k)}) \geq 0 \), then the 1-symbol update search terminates. The data vector at this point is referred to as ‘1-symbol update local minima.’ After reaching the 1-symbol update local minima, we look for a further decrease in the cost function by updating multiple symbols simultaneously.
2.2.2 Why Multiple Symbol Updates?

The motivation for trying out multiple symbol updates can be explained as follows. The following discussion is for non-orthogonal STBC MIMO systems with \(2N_t^2\) real symbols, and can be easily extended to V-BLAST MIMO systems with \(2N_t^2\) real symbols. Let \(\mathcal{L}_K \subseteq \mathcal{S}\) denote the set of data vectors such that for any \(d \in \mathcal{L}_K\), if a \(K\)-symbol update is performed on \(d\) resulting in a vector \(d'\), then \(||y - Hd'|| \geq ||y - Hd||\). We note that \(d_{ML} \in \mathcal{L}_K\), \(\forall K = 1, 2, \ldots, 2N_t^2\), because any number of symbol updates on \(d_{ML}\) will not decrease the cost function. We define another set \(\mathcal{M}_K = \bigcap_{j=1}^{K} \mathcal{L}_j\). Note that \(d_{ML} \in \mathcal{M}_K\), \(\forall K = 1, 2, \ldots, 2N_t^2\), and \(\mathcal{M}_{2N_t^2} = \{d_{ML}\}\), i.e., \(\mathcal{M}_{2N_t^2}\) is a singleton set with \(d_{ML}\) as the only element. It is noted that if the updates are done optimally, then the output of the \(K\)-LAS algorithm converges to a vector in \(\mathcal{M}_K\). Also, \(|\mathcal{M}_{K+1}| \leq |\mathcal{M}_K|, K = 1, 2, \ldots, 2N_t^2 - 1\). For any \(d \in \mathcal{M}_K\), \(K = 1, 2, \ldots, 2N_t^2\) and \(d \neq d_{ML}\), it can be seen that \(d\) and \(d_{ML}\) will differ in \(K + 1\) or more locations. At high SNR, \(d_{ML} = x\) with high probability, and therefore with high probability, the separation between \(d \in \mathcal{M}_K\) and \(x = d_{ML}\) will monotonically increase with increasing \(K\). Since \(d_{ML} \in \mathcal{M}_K\), and \(|\mathcal{M}_K|\) decreases monotonically with increasing \(K\), there will be lesser non-ML data vectors to which the algorithm can converge to for increasing \(K\). With increasing \(K\), due to increased separation between \(d \in \mathcal{M}_K\) and \(x\), and also lesser number of non-ML vectors in \(\mathcal{M}_K\), the probability of the noise vector \(n\) inducing an error would decrease. This indicates that \(K\)-symbol updates with large \(K\) could get near to ML performance with increasing complexity for increasing \(K^1\).

2.2.3 \(K\)-symbol Update

We continue our description of the algorithm for the case of non-orthogonal STBC MIMO systems with \(2N_t^2\) real symbols. A similar description holds for V-BLAST MIMO

\(^1\)In Chapter 3, through an asymptotic performance analysis, we will try to get more insights into the performance of the LAS detection algorithm.
with $2N_t$ real symbols.

In this subsection, we present the update algorithm for the general case where $K$ symbols, $1 < K \leq 2N_t^2$, are updated simultaneously in one iteration. $K$-symbol updates can be done in $\binom{2N_t^2}{K}$ ways, among which we seek to find that update which gives the largest reduction in the ML cost. Assume that in the $(k+1)$th iteration, $K$ symbols at the indices $i_1, i_2, \ldots, i_K$ of $d^{(k)}$ are updated. Each $i_j, j = 1, 2, \ldots, K$, can take values from $1, 2, \ldots, N_t^2$ for $M$-PAM and $1, 2, \ldots, 2N_t^2$ for $M$-QAM. Further, define the set of indices, $\mathcal{U} = \{i_1, i_2, \ldots, i_K\}$. The update rule for the $K$-symbol update can then be written as

$$d^{(k+1)} = d^{(k)} + \sum_{j=1}^{K} \lambda_{ij}^{(k)} e_{ij}, \quad (2.32)$$

For any iteration $k$, $d^{(k)}$ belongs to the space $\mathcal{S}$, and therefore $\lambda_{ij}^{(k)}$ can take only certain integer values. In particular, $\lambda_{ij}^{(k)} \in \mathcal{A}_{ij}^{(k)}$, where $\mathcal{A}_{ij}^{(k)} \triangleq \{ x | (x + d_{ij}^{(k)}) \in \mathcal{A}_{ij}, x \neq 0 \}$.

For example, for 16-QAM, $\mathcal{A}_{ij} = \{-3, -1, 1, 3\}$, and if $d_{ij}^{(k)}$ is -1, then $\mathcal{A}_{ij}^{(k)} = \{-2, 2, 4\}$.

Using (2.18), we can write the cost difference function $\Delta C_{\mathcal{U}}^{(k+1)}(\lambda_{i_1}^{(k)}, \lambda_{i_2}^{(k)}, \ldots, \lambda_{i_K}^{(k)}) \triangleq C^{(k+1)} - C^{(k)}$ as

$$\Delta C_{\mathcal{U}}^{(k+1)}(\lambda_{i_1}^{(k)}, \lambda_{i_2}^{(k)}, \ldots, \lambda_{i_K}^{(k)}) = \sum_{j=1}^{K} \lambda_{ij}^{(k)} (G)_{ij}, i_j$$

$$+ 2 \sum_{q=1}^{K} \sum_{p=q+1}^{K} \lambda_{ip}^{(k)} \lambda_{iq}^{(k)} (G)_{ip}, iq - 2 \sum_{j=1}^{K} \lambda_{ij}^{(k)} \sim_{ij}, \quad (2.33)$$

where $\lambda_{ij}^{(k)} \in \mathcal{A}_{ij}^{(k)}$, which can be compactly written as $(\lambda_{i_1}^{(k)}, \lambda_{i_2}^{(k)}, \ldots, \lambda_{i_K}^{(k)}) \in \mathcal{A}_{\mathcal{U}}^{(k)}$, where $\mathcal{A}_{\mathcal{U}}^{(k)}$ denotes the Cartesian product of $\mathcal{A}_{i_1}^{(k)}, \mathcal{A}_{i_2}^{(k)}$ through to $\mathcal{A}_{i_K}^{(k)}$.

For a given $\mathcal{U}$, in order to decrease the ML cost, we would like to choose the value of the $K$-tuple $(\lambda_{i_1}^{(k)}, \lambda_{i_2}^{(k)}, \ldots, \lambda_{i_K}^{(k)})$ such that the cost difference given by (2.33) is negative. If multiple $K$-tuples exist for which the cost difference is negative, we choose the $K$-tuple which gives the most negative cost difference.
Unlike for 1-symbol update, for $K$-symbol update we do not have a closed-form expression for $(\lambda_{i_1, opt}^{(k)}, \lambda_{i_2, opt}^{(k)}, \ldots, \lambda_{i_K, opt}^{(k)})$ which minimizes the cost difference over $\Lambda_{it}^{(k)}$, since the cost difference is a function of $K$ discrete valued variables. Consequently, a brute-force method is to evaluate $\Delta C_{it}^{k+1}(\lambda_{i_1}^{(k)}, \lambda_{i_2}^{(k)}, \ldots, \lambda_{i_K}^{(k)})$ over all possible values of $(\lambda_{i_1}^{(k)}, \lambda_{i_2}^{(k)}, \ldots, \lambda_{i_K}^{(k)})$. Approximate methods can be adopted to solve this problem using lesser complexity. One method based on zero-forcing is as follows. The cost difference function in (2.33) can be rewritten as

$$
\Delta C_{it}^{k+1}(\lambda_{i_1}^{(k)}, \lambda_{i_2}^{(k)}, \ldots, \lambda_{i_K}^{(k)}) = \Lambda_{it}^{(k)} F_{it} \Lambda_{it}^{(k)} - 2\Lambda_{it}^{(k)T} z_{it}^{(k)},
$$

(2.34)

where $\Lambda_{it}^{(k)} \triangleq [\lambda_{i_1}^{(k)} \lambda_{i_2}^{(k)} \cdots \lambda_{i_K}^{(k)}]^T$, $z_{it}^{(k)} \triangleq [z_{i_1}^{(k)} z_{i_2}^{(k)} \cdots z_{i_K}^{(k)}]^T$, and $F_{it} \in \mathbb{R}^{K \times K}$, where $(F_{it})_{p,q} = (G)_{ip, iq}$ and $p, q \in \{1, 2, \cdots, K\}$. Since $\Delta C_{it}^{k+1}(\lambda_{i_1}^{(k)}, \lambda_{i_2}^{(k)}, \ldots, \lambda_{i_K}^{(k)})$ is a strictly convex quadratic function of $\Lambda_{it}^{(k)}$ (the Hessian $F_{it}$ is positive definite with probability 1), a unique global minima exists, and is given by

$$
\hat{\Lambda}_{it}^{(k)} = F_{it}^{-1} z_{it}^{(k)}.
$$

(2.35)

However, the solution given by (2.35) need not lie in $\Lambda_{it}^{(k)}$. So, we first round-off the solution as

$$
\tilde{\Lambda}_{it}^{(k)} = 2 \left[ 0.5 \hat{\Lambda}_{it}^{(k)} \right],
$$

(2.36)

where the operation in (2.36) is done element-wise, since $\hat{\Lambda}_{it}^{(k)}$ is a vector. Further, let $\tilde{\Lambda}_{it}^{(k)} \triangleq [\tilde{\lambda}_{i_1}^{(k)} \tilde{\lambda}_{i_2}^{(k)} \cdots \tilde{\lambda}_{i_K}^{(k)}]^T$. It is still possible that the solution $\tilde{\Lambda}_{it}^{(k)}$ in (2.36) need not lie in $\Lambda_{it}^{(k)}$. This would result in $d_{ij}^{(k+1)} \notin \Lambda_{ij}$ for some $j$. For example, if $\Lambda_{ij}$ is $M$-PAM, then $d_{ij}^{(k+1)} \notin \Lambda_{ij}$ if $d_{ij}^{(k)} + \tilde{\lambda}_{ij}^{(k)} > (M-1)$ or $d_{ij}^{(k)} + \tilde{\lambda}_{ij}^{(k)} < -(M-1)$. In such cases, we propose the following adjustment to $\tilde{\lambda}_{ij}^{(k)}$ for $j = 1, 2, \cdots, K$:...
\[
\hat{\lambda}^{(k)}_{i_j} = \begin{cases} 
(M - 1) - d^{(k)}_{i_j}, & \text{when } \hat{\lambda}^{(k)}_{i_j} + d^{(k)}_{i_j} > (M - 1) \\
-(M - 1) - d^{(k)}_{i_j}, & \text{when } \hat{\lambda}^{(k)}_{i_j} + d^{(k)}_{i_j} < -(M - 1) \end{cases} 
\] (2.37)

After these adjustments, we are guaranteed that \(\hat{\Lambda}^{(k)} \in A^{(k)}_U\). Therefore, the new cost difference function value is given by:

\[
\Delta C^{k+1}_{\hat{U}}(\hat{\Lambda}^{(k)}_{i_1}, \hat{\Lambda}^{(k)}_{i_2}, \ldots, \hat{\Lambda}^{(k)}_{i_K}).
\] (2.38)

The \(K\)-update is successful and the update is done only if \(\Delta C^{k+1}_{\hat{U}}(\hat{\Lambda}^{(k)}_{i_1}, \hat{\Lambda}^{(k)}_{i_2}, \ldots, \hat{\Lambda}^{(k)}_{i_K}) < 0\).

The update rules for the \(z^{(k)}\) and \(d^{(k)}\) vectors are given by:

\[
z^{(k+1)} = z^{(k)} - \sum_{j=1}^{K} \hat{\lambda}^{(k)}_{i_j} g_{i_j},
\] (2.39)

\[
d^{(k+1)} = d^{(k)} + \sum_{j=1}^{K} \hat{\lambda}^{(k)}_{i_j} e_{i_j}.
\] (2.40)

### 2.2.4 Computational Complexity of the \(M\)-LAS Algorithm

The complexity of the \(M\)-LAS algorithm comprises of three components, namely, 

1) computation of the initial vector \(d^{(0)}\),

2) computation of \(H^T H\), and

3) the search operation.

**Complexity for V-BLAST MIMO:** For V-BLAST MIMO systems, the combined complexity of computing \(d^{(0)}\) and \(H^T H\) is \(O(N_t^2 N_r) + O(N_t^3)\). Since \(N_r \geq N_t\), this complexity is simply \(O(N_t^2 N_r)\). As shown in Fig. 2.1, the mean number of LAS stages for the 3-LAS...
detector is found to be constant (we found it to be true for 2-LAS also). The mean number of iterations in each stage is, however, proportional to $N_t$, as is illustrated in Fig. 2.2. The first sub-stage of each stage consists of a sequence of one update iterations till the algorithm reaches a 1-update local minima. The complexity of each iteration is $O(N_r)$, and since the mean number of such iterations per stage is $O(N_t)$, the total complexity of the first sub-stage alone is $O(N_t N_r)$. For the 1-LAS algorithm, since there is only the first sub-stage in the only stage, the total search complexity is $O(N_t N_r)$. For the 2-LAS detector, the complexity of the second sub-stage is $O(N_t^2)$ (since all possible 2-symbol updates are tried in order to reduce the ML cost). Since the number of stages is constant, it follows that the total search complexity of 2-LAS is $O(N_t^2) + O(N_t N_r)$. Similarly, the complexity of the 3-LAS detector is $O(N_t^3) + O(N_t N_r)$. In general, the total complexity of an $M$-LAS detector is $O(N_t^M) + O(N_t N_r)$. Upon adding the complexity of the initial vector and $H^T H$, it can be concluded that the total complexity of 1-, 2-, and 3-LAS detectors is $O(N_t^2 N_r)$. Since there are $N_t$ symbols per transmitted vector, the average per-symbol complexity is $O(N_t N_r)$. 

Figure 2.1: Mean number of stages for the 3-LAS algorithm with MMSE initial vector. V-BLAST MIMO, 4-QAM, SNR = 6, 9, 12 dB.
Chapter 2. Large-MIMO Detection Using Likelihood Ascent Search

Figure 2.2: Mean number of iterations per stage per transmit antenna for the 3-LAS algorithm with MMSE initial vector. V-BLAST MIMO, 4-QAM, SNR = 6, 9, 12 dB.

Complexity for Non-Orthogonal STBC MIMO: Next, we discuss the computational complexity of the $M$-LAS algorithm, when used to detect non-orthogonal STBC MIMO signals. Figure 2.3 shows the per-symbol complexity plots as a function of $N_t = N_r$ for 4-QAM at an SNR of 6 dB using MMSE initial vector. Two good properties of the STBCs from CDA are useful in achieving low orders of complexity for the computation of $d^{(0)}$ and $H^T H$. They are: i) the weight matrices $A_c^{(i)}$’s are permutation type, and ii) the $N_t^2 \times N_t^2$ matrix formed with $N_t^2 \times 1$-sized $a_c^{(i)}$ vectors as columns is a scaled unitary matrix. These properties allow the computation of MMSE/ZF initial solution in $O(N_t^3 N_r)$ complexity, i.e., in $O(N_t N_r)$ per-symbol complexity since there are $N_t^2$ symbols in one STBC matrix. Likewise, the computation of $H^T H$ can be done in $O(N_t^3)$ per-symbol complexity.

The average per-symbol complexities of the 1-LAS and 2-LAS search operations are $O(N_t^2)$ and $O(N_t^2 \log N_t)$, respectively, which can be explained as follows. The average search complexity is the complexity of one search stage times the mean number of search stages till the algorithm terminates. For 1-LAS, the number of search stages is
always one. There are multiple iterations in the search, and in each iteration all possible \( (2N_t^2) \) 1-symbol updates are considered. So, the per-iteration complexity in 1-LAS is \( O(N_t^2) \), i.e., \( O(1) \) complexity per symbol. Further, the mean number of iterations before the algorithm terminates in 1-LAS was found to be \( O(N_t^2) \) through simulations. So, the overall per-symbol search complexity of 1-LAS is \( O(N_t^2) \). In 2-LAS, the complexity of the 2-symbol update dominates over the 1-symbol update. Since there are \( (2N_t^2) \) possible 2-symbol updates, the complexity of one search stage is \( O(N_t^4) \), i.e., \( O(N_t^2) \) complexity per symbol. The mean number of stages till the algorithm terminates in 2-LAS was found to be \( O(\log N_t) \) through simulations. Therefore, the overall per-symbol search complexity of 2-LAS is \( O(N_t^2 \log N_t) \). These can be observed from Fig. 2.3, where it can be seen that the per-symbol complexity in the initial vector computation plus the 1-LAS/2-LAS search operation is \( O(N_t^2) / O(N_t^2 \log N_t) \); i.e., 1-LAS and 2-LAS complexity plots run parallel to the \( c_1 N_t^2 \) and \( c_2 N_t^2 \log N_t \) lines, respectively. With the computation of \( H^T H \) included, the complexity order is more than \( N_t^2 \). From the slopes of the plots in Fig. 2.3, we find that the overall complexities for \( N_t = 16 \) and 32 are proportional to \( N_t^{2.5} \) and \( N_t^{2.7} \), respectively.

**Further Complexity Reduction for ILL-Only STBC:** For the special case of ILL-only STBCs (i.e., \( \delta = t = 1 \)), the complexity involved in computing \( d^{(0)} \) and \( H^T H \) can be reduced further. This becomes possible due to the following property of ILL-only STBCs. Let \( V_a \) be the complex \( N_t^2 \times N_t^2 \) matrix with \( a_c^{(i)} \) as its \( i \)th column. The computation of \( d^{(0)} \) (or \( H^T H \)) involves multiplication of \( V_a^H \) with another vector (or matrix). The columns of \( V_a^H \) can be permuted in such a way that the permuted matrix is block-diagonal, where each block is a \( N_t \times N_t \) DFT matrix for \( \delta = t = 1 \). So, the multiplication of \( V_a^H \) by any vector becomes equivalent to a \( N_t \)-point DFT operation, which can be efficiently computed using FFT in \( O(N_t \log N_t) \) complexity. Using this simplification, the per-symbol complexity of computing \( H^T H \) is reduced from \( O(N_t^3) \) to \( O(N_t^2 \log N_t) \). Computing \( d^{(0)} \) using MMSE filter involves the computation of \( \frac{1}{N_t} V_a^H (I \otimes ((H_c^H H_c + \ldots \right) \)
\[ \frac{1}{\gamma N_t} \mathbf{I}^{-1} \mathbf{H}_c^H \mathbf{y}_c. \] The complexity of computing the vector \( (\mathbf{I} \otimes ((\mathbf{H}_c^H \mathbf{H}_c + \frac{1}{\gamma N_t} \mathbf{I})^{-1} \mathbf{H}_c^H)) \mathbf{y}_c \) is \( O(N_r^2 N_t) \), and the complexity of computing \( \mathbf{V}_a^H (\mathbf{I} \otimes ((\mathbf{H}_c^H \mathbf{H}_c + \frac{1}{\gamma N_t} \mathbf{I})^{-1} \mathbf{H}_c^H)) \mathbf{y}_c \) is \( O(N_t^3 N_r) \). In the case of ILL-only STBC, because of the above-mentioned property, the complexity of computing \( \mathbf{V}_a^H (\mathbf{I} \otimes ((\mathbf{H}_c^H \mathbf{H}_c + \frac{1}{\gamma N_t} \mathbf{I})^{-1} \mathbf{H}_c^H)) \mathbf{y}_c \) gets reduced to \( O(N_t^2 \log N_t) \) from \( O(N_t^3 N_r) \). So the total complexity for computing \( d^{(0)} \) in ILL-only STBC is \( O(N_t^2 N_r) + O(N_t^2 \log N_t) \), which gives a per-symbol complexity of \( O(N_r) + O(\log N_t) \). So, the overall per-symbol complexity for 1-LAS detection of ILL-STBCs is \( O(N_t^2 \log N_t) \).

### 2.2.5 Generation of Soft Outputs

We propose to generate soft values at the \( M \)-LAS output for all the individual bits that constitute the \( M \)-PAM/M-QAM symbols as follows. The method is described for STBC MIMO, but is applicable for V-BLAST MIMO as well. These output values are fed as soft inputs to the decoder in a coded system. Let \( \mathbf{d} = [\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_{2N_t}] \), \( \hat{x}_i \in A_i \) denote the detected output symbol vector from the \( M \)-LAS algorithm. Let the symbol \( \hat{x}_i \) map to the bit vector \( \mathbf{b}_i = [b_{i,1}, b_{i,2}, \ldots, b_{i,K_i}]^T \), where \( K_i = \log_2 |A_i| \), and \( b_{i,j} \in \{+1, -1\} \),...
Let \( i = 1, 2, \cdots, 2N_t^2 \) and \( j = 1, 2, \cdots, K_i \). Let \( \tilde{b}_{i,j} \in \mathbb{R} \) denote the soft value for the \( j \)th bit of the \( i \)th symbol. Given \( \mathbf{d} \), we need to find \( \tilde{b}_{i,j} \), \( \forall (i,j) \).

Note that the quantity \( \| \mathbf{y} - \mathbf{Hd} \|_2^2 \) is inversely related to the likelihood that \( \mathbf{d} \) is indeed the transmitted symbol vector. Let the \( \mathbf{d} \) vector with its \( j \)th bit of the \( i \)th symbol forced to +1 be denoted as vector \( \mathbf{d}_{i}^{j,+} \). Likewise, let \( \mathbf{d}_{i}^{j,-} \) be the vector \( \mathbf{d} \) with its \( j \)th bit of the \( i \)th symbol forced to -1. Then the quantities \( \| \mathbf{y} - \mathbf{Hd}_{i}^{j,+} \|_2^2 \) and \( \| \mathbf{y} - \mathbf{Hd}_{i}^{j,-} \|_2^2 \) are inversely related to the likelihoods that the \( j \)th bit of the \( i \)th transmitted symbol is +1 and -1, respectively. So, if \( \| \mathbf{y} - \mathbf{Hd}_{i}^{j,-} \|_2^2 - \| \mathbf{y} - \mathbf{Hd}_{i}^{j,+} \|_2^2 \) is +ve (or -ve), it indicates that the \( j \)th bit of the \( i \)th transmitted symbol has a higher likelihood of being +1 (or -1).

So, the quantity \( \| \mathbf{y} - \mathbf{Hd}_{i}^{j,-} \|_2^2 - \| \mathbf{y} - \mathbf{Hd}_{i}^{j,+} \|_2^2 \), appropriately normalized to avoid unbounded increase for increasing \( N_t \), can be a good soft value for the \( j \)th bit of the \( i \)th symbol.

With this motivation, we generate the soft output value for the \( j \)th bit of the \( i \)th symbol as

\[
\tilde{b}_{i,j} = \frac{\| \mathbf{y} - \mathbf{Hd}_{i}^{j,-} \|_2^2 - \| \mathbf{y} - \mathbf{Hd}_{i}^{j,+} \|_2^2}{\| \mathbf{h}_i \|_2^2}, \tag{2.41}
\]

where the normalization by \( \| \mathbf{h}_i \|_2^2 \) is to contain unbounded increase of \( \tilde{b}_{i,j} \) for increasing \( N_t \). The RHS in the above can be efficiently computed in terms of \( z \) and \( G \) as follows.

Since \( \mathbf{d}_{i}^{j,+} \) and \( \mathbf{d}_{i}^{j,-} \) differ only in the \( i \)th entry, we can write

\[
\mathbf{d}_{i}^{j,-} = \mathbf{d}_{i}^{j,+} + \lambda_{i,j} \mathbf{e}_i. \tag{2.42}
\]

Since we know \( \mathbf{d}_{i}^{j,-} \) and \( \mathbf{d}_{i}^{j,+} \), we know \( \lambda_{i,j} \) from (2.42). Substituting (2.42) in (2.41), we can write

\[
\tilde{b}_{i,j} \| \mathbf{h}_i \|_2^2 = \| \mathbf{y} - \mathbf{Hd}_{i}^{j,+} - \lambda_{i,j} \mathbf{h}_i \|_2^2 - \| \mathbf{y} - \mathbf{Hd}_{i}^{j,+} \|_2^2
\]
\[
= \lambda_{i,j}^2 \| \mathbf{h}_i \|_2^2 - 2\lambda_{i,j} \mathbf{h}_i^T (\mathbf{y} - \mathbf{Hd}_{i}^{j,+}) \tag{2.43}
\]
\[
= -\lambda_{i,j}^2 \| \mathbf{h}_i \|_2^2 - 2\lambda_{i,j} \mathbf{h}_i^T (\mathbf{y} - \mathbf{Hd}_{i}^{j,-}). \tag{2.44}
\]
If $b_{i,j} = 1$, then $d_i^+ = d$ and substituting this in (2.43) and dividing by $\|h_i\|^2$, we get

$$\tilde{b}_{i,j} = \lambda_{i,j}^2 - 2\lambda_{i,j} \frac{z_i}{(G)_{i,i}}.$$  

(2.45)

If $b_{i,j} = -1$, then $d_i^- = d$ and substituting this in (2.44) and dividing by $\|h_i\|^2$, we get

$$\tilde{b}_{i,j} = -\lambda_{i,j}^2 - 2\lambda_{i,j} \frac{z_i}{(G)_{i,i}}.$$  

(2.46)

It is noted that $z$ and $G$ are already available upon the termination of the $M$-LAS algorithm, and hence the complexity of computing $\tilde{b}_{i,j}$ in (2.45) and (2.46) is constant. Hence, the overall complexity in computing the soft values for all the bits is $O(N_t \log_2 M)$. We also see from (2.45) and (2.46) that the magnitude of $\tilde{b}_{i,j}$ depends upon $\lambda_{i,j}$. For large-size signal sets, the possible values of $\lambda_{i,j}$ will also be large in magnitude. We therefore have to normalize $\tilde{b}_{i,j}$ for the turbo decoder to function properly. It has been observed through simulations that normalizing $\tilde{b}_{i,j}$ by $(\frac{\lambda_{i,j}}{2})^2$ resulted in good performance.

### 2.3 Performance in Large V-BLAST MIMO

In this section, we report uncoded and coded BER performance of the $M$-LAS detector for V-BLAST MIMO systems.

*Uncoded BER Performance:* In Fig. 2.4, we present the uncoded BER of 3-LAS detector for different values of $N_t = N_r$ and 4-QAM, obtained through simulations. MMSE filter is used as the initial filter. MMSE filter performance without subsequent LAS search is also plotted for comparison. As we mentioned earlier, we use the BER performance in SISO unfaded AWGN channel as a lower bound for comparison. It is observed that the BER performance of 3-LAS improves with increasing $N_r = N_t$. For e.g., at an uncoded BER of $10^{-3}$, with $N_r = N_t = 16$, the performance of 3-LAS is about 3 dB away from SISO AWGN performance, whereas with $N_r = N_t = 256$, the performance gets
close to within just 0.4 dB from the SISO AWGN performance. We refer to this behavior as ‘large-system behavior’ of the LAS algorithm. We observed that MMSE-only performance without subsequent LAS search do not improve for increasing $N_t = N_r$. Therefore, the proposed LAS search is found to be an effective mechanism to improve upon the performance of the initial MMSE solution at the same order of complexity as that of the MMSE solution.

In Fig. 2.5, we plot the SNR required to achieve a target BER of $10^{-3}$ with 1-LAS detector for increasing $N_t = N_r$ and 4-QAM. MMSE initial vector is used. It is observed that, the SNR required to meet the target BER indeed gets closer to the SISO AWGN SNR for increasing $N_t = N_r$. It is interesting to see that 1-LAS performance is far from SISO AWGN performance for small values of $N_r = N_t$. Therefore, in a way, large dimensions are actually favorable for the convergence of the LAS algorithm to the correct transmitted vector. A similar large-system behavior is observed with 16-QAM as well, which is illustrated in Fig. 2.6.

Next, in Fig. 2.7, we compare the performance of 1-LAS and 3-LAS detection for $N_t = N_r = 64, 32$ and 4-QAM. It can be seen that 3-LAS outperforms 1-LAS. This
Chapter 2. Large-MIMO Detection Using Likelihood Ascent Search

Figure 2.5: Average received SNR required to achieve a target BER of $10^{-3}$ in V-BLAST MIMO for increasing values of $N_t = N_r$ using 1-LAS detection with MMSE initial vector. 4-QAM.

Figure 2.6: Average received SNR required to achieve a target BER of $10^{-4}$ in V-BLAST MIMO for increasing values of $N_t = N_r$ using 1-LAS detection with MMSE initial vector. 16-QAM.
performance improvement is due to the 2- and 3-symbol updates performed in 3-LAS, in addition to the 1-symbol updates performed in 1-LAS. As pointed out earlier, the 2- and 3-symbol updates in 3-LAS increase the complexity a little, but the average per-symbol complexity still remains as $O(N_tN_r)$.

**Turbo Coded BER Performance:** Figure 2.8 shows the rate-3/4 turbo coded BER performance of 3-LAS detection with MMSE initial filter for $N_t = N_r = 64, 128$ and 4-QAM. We have also shown the minimum SNRs required to achieve theoretical capacity in $64 \times 64$ and $128 \times 128$ MIMO channels with perfect CSIR, evaluated using the MIMO capacity formula given by [2]

\[
C = \mathbb{E} \left[ \log \det \left( \mathbf{I}_{N_r} + \left( \frac{\gamma}{N_t} \right) \mathbf{HH}^H \right) \right], \tag{2.47}
\]

where $\gamma$ is the average SNR per receive antenna. Performance with hard decision and soft decision inputs to the turbo decoder are plotted. Soft outputs generated using the method described in Sec. 2.2.5 are fed as soft inputs to the turbo decoder. From Fig. 2.8, we observe that

i) with soft inputs to the turbo decoder, the performance improves by about 1 dB compared to that with hard decision inputs, and

ii) 3-LAS detection
Figure 2.8: Turbo coded BER performance of 3-LAS detection for $N_t = N_r = 64, 128$, 4-QAM, and rate-3/4 turbo code. MMSE initial vector.

achieves a coded performance (i.e., vertical fall in coded BER) which is close to within about 4.5 dB from theoretical capacity.

2.4 Performance in Large Non-orthogonal STBC MIMO

In this section, we present the uncoded/turbo coded BER performance of the $M$-LAS detector in decoding non-orthogonal STBCs from CDA, assuming perfect knowledge of CSIR. In all the BER simulations in this section, we have assumed that the fade remains constant over one STBC matrix duration and varies i.i.d. from one STBC matrix duration to the other. We consider two STBC designs; i) ‘FD-ILL’ STBCs where $\delta = e^{\sqrt{5}j}, t = e^j$ in (2.15), and ii) ‘ILL-only’ STBCs where $\delta = t = 1$. The SNRs in all the BER performance figures are the average received SNR per received antenna, $\gamma$, defined in Section 2.1.4 [2]. We have used MMSE filter as the initial filter in all the simulations.

---

$^2$We will relax this perfect channel knowledge assumption in the next section, where we present an iterative detection/channel estimation scheme for the considered large STBC MIMO system.
Figure 2.9: Uncoded BER performance of 1-LAS, 2-LAS and 3-LAS detectors for ILL-only STBCs for different $N_t = N_r$. 4-QAM, $2N_t$ bps/Hz.

### 2.4.1 Uncoded BER as a Function of Increasing $N_t = N_r$

In Fig. 2.9, we plot the uncoded BER performance of the 1-, 2-, and 3-LAS algorithms in decoding ILL-only STBCs ($4 \times 4$, $8 \times 8$, $16 \times 16$, $32 \times 32$ STBCs) for $N_t = N_r = 4, 8, 16, 32$ and 4-QAM. SISO AWGN performance (without fading) and MMSE-only performance (i.e., without the search using LAS) are also plotted for comparison. It can be seen that MMSE-only performance does not improve with increasing STBC size (i.e., increasing $N_t = N_r$). However, it is interesting to see that, when the proposed search using LAS is performed following the MMSE operation, the performance improves for increasing $N_t = N_r$, illustrating the performance benefit due to the proposed search strategy. For example, though the LAS detector performs far from SISO AWGN performance for small number of dimensions (e.g., $4 \times 4, 8 \times 8$ STBCs with 32 and 128 real dimensions, respectively), its large system behavior at increased number of dimensions (e.g., $16 \times 16$ and $32 \times 32$ STBCs with 512 and 2048 real dimensions, respectively) effectively renders near SISO AWGN performance; e.g., with $N_t = N_r = 16, 32$, for BERs better than $10^{-3}$, the LAS detector performs very close to SISO AWGN performance. We also
observe that 3-LAS performs better than 2-LAS for $N_t = N_r = 4$, 8, and 2-LAS performs better than 1-LAS. Since close to SISO AWGN performance is achieved with 1-, 2-, or 3-symbol update itself, the cases of more than 3-symbol update, which will result in increased complexity with diminishing returns in performance gain, are not considered in the performance evaluation.

### 2.4.2 Performance of FD-ILL versus ILL-only STBCs

In Fig. 2.10, we present a uncoded BER performance comparison between FD-ILL versus ILL-only STBCs for 4-QAM at different $N_t = N_r$ using 1-LAS detection. The BER plots in Fig. 2.10 illustrate that the performance of ILL-only STBCs with 1-LAS detection for $N_t = N_r = 4, 8, 16, 32$ and 4-QAM are almost as good as those of the corresponding FD-ILL STBCs. A similar closeness between the performance of ILL-only and FD-ILL STBCs is observed in the turbo coded BER performance as well, which is shown in Fig. 2.15 for a $16 \times 16$ STBC with 4-QAM and turbo code rates of 1/3, 1/2 and 3/4. This is an interesting observation, since this suggests that, in such cases, the
computational complexity advantage with $\delta = t = 1$ in ILL-only STBCs can be taken advantage of without incurring much performance loss compared to FD-ILL STBCs.

2.4.3 Decoding and BER of Perfect Codes of Large Dimensions

While the STBC design in (2.15) offers both ILL and FD, perfect codes\(^3\) under ML decoding can provide coding gain in addition to ILL and FD [84]-[22]. Decoding of perfect codes has been reported in the literature for only up to 5 antennas using sphere/lattice decoding [87]. The complexity of these decoders are prohibitive for decoding large-sized perfect codes, although large-sized codes are of interest from a high spectral efficiency view point. We note that, because of its low-complexity attribute, the $M$-LAS detector is able to decode perfect codes of large dimensions. In Figs. 2.11 and 2.12, we present the simulated BER performance of perfect codes in comparison with those of ILL-only and FD-ILL STBCs for up to 32 transmit antennas using 1-LAS detector.

In Fig. 2.11, we show uncoded BER comparison between perfect codes and ILL-only STBCs for different $N_t = N_r$ and 4-QAM using 1-LAS detection. The $4 \times 4$ and $6 \times 6$ perfect codes are from [86], and the $8 \times 8$, $16 \times 16$ and $32 \times 32$ perfect codes are from [87]. From Fig. 2.11, it can be seen that the 1-LAS detector achieves better performance for ILL-only STBCs than for perfect codes, when codes with small number of transmit antennas are considered (e.g., $N_t = 4, 6, 8$). While perfect codes are expected to perform better than ILL-only codes under ML detection for any $N_t$, we observe the opposite behavior under 1-LAS detection for small $N_t$ (i.e., ILL-only STBCs performing better than perfect codes for small dimensions). This behavior could be attributed to the nature of the LAS detector, which achieves near-optimal performance only when the number of dimensions is large, and it appears that, in the detection process, LAS is more effective in disentangling the symbols in STBCs when $\delta = t = 1$ (i.e., in ILL-only STBCs) than in

\(^3\)We note that the definition of perfect codes differ in [86] and [87]. The perfect codes covered by the definition in [87] includes the perfect codes of [86] as a proper subclass. However, for our purpose of illustrating the performance of the proposed detector in large STBC MIMO systems, we refer to the codes in [86] as well as [87] as perfect codes.
perfect codes. The performance gap between perfect codes and ILL-only STBCs with 1-LAS detection diminishes for increasing code sizes such that the performance for $32 \times 32$ perfect code and ILL-only STBC with 4-QAM are almost same and close to the SISO AWGN performance. In Fig. 2.12, we show a similar comparison between perfect codes, ILL-only and FD-ILL only STBCs when larger modulation alphabet sizes (e.g., 16-QAM) are used in the case of $16 \times 16$ and $32 \times 32$ codes. It can be seen that with higher-order QAM like 16-QAM, perfect codes with 1-LAS detection perform poorer than ILL-only and FD-ILL STBCs, and that ILL-only and FD-ILL STBCs perform almost same and close to the SISO AWGN performance. The results in Figs. 2.11 and 2.12 suggest that, with 1-LAS detection, owing to the complexity advantage and good performance in using $\delta = t = 1$, ILL-only STBCs can be a good choice for practical large STBC MIMO systems [109],[110].
Figure 2.12: Uncoded BER performance comparison between perfect codes, ILL-only, and FD-ILL STBCs for $N_t = N_r = 16, 32, 16$-QAM, $4N_t$ bps/Hz, 1-LAS detection.

### 2.4.4 Comparison with Other Large-MIMO Architecture/Detector Combinations

In [111], Choi et al have presented an iterative soft interference cancellation (ISIC) scheme for multiple antenna systems, derived based on maximum a posteriori (MAP) criterion. We compared the performance of the ISIC scheme in [111] with that of the 1-LAS algorithm in detecting $4 \times 4$, $8 \times 8$ and $16 \times 16$ ILL-only STBCs with $N_t = N_r$ and 4-QAM. Figure 2.13 shows this performance comparison. In [111], zero-forcing vector was used as the initial vector in the ISIC scheme. However, performance is better with MMSE initial vector. Since we used MMSE initial vector for 1-LAS, we have used MMSE initial vector for the ISIC algorithm as well. Also, in [111], 4 to 5 iterations were shown to be good enough for the ISIC algorithm to converge. In our simulations of the ISIC algorithm, we used 10 iterations. Two key observations can be made from Fig. 2.13: 1) like the 1-LAS algorithm, the ISIC algorithm also shows large system behavior (i.e., improved BER for increasing $N_t = N_r$), and 2) the 1-LAS algorithm outperforms the ISIC algorithm by about 3 to 5 dB at $10^{-3}$ uncoded BER. In addition, the complexity
of the ISIC scheme is higher than the proposed scheme (see the complexity comparison in Table 2.1).

Next, we compare the proposed large-MIMO architecture using STBCs from CDA and $M$-LAS detection with other large-MIMO architectures and associated detectors reported in the literature. Large-MIMO architectures that use stacking of multiple small-sized STBCs and interference cancellation (IC) detectors for these schemes have been investigated in [88],[112],[113]. Here, we compare different architecture/detector combinations, fixing the total number of transmit/receive antennas and spectral efficiency to be same in all the considered combinations. Specifically, we fix $N_t = N_r = 16$ and a spectral efficiency of 32 bps/Hz for all the combinations. We compare the following seven different architecture/detector combinations which use the same $N_t = N_r = 16$ and achieve 32 bps/Hz spectral efficiency (see Table 2.1): i) proposed scheme using $16 \times 16$ ILL-only STBC (rate-16) with 4-QAM and 1-LAS detection, ii) $16 \times 16$ ILL-only STBC (rate-16) with 4-QAM and ISIC algorithm in [111] with 10 iterations, iii) four $4 \times 4$ stacked QOSTBCs (rate-1) with 256-QAM and IC algorithm presented in [88], iv)
Figure 2.14: Uncoded BER comparison between different large-MIMO architecture/detector combinations for given number of transmit/receive antennas ($N_t = N_r = 16$) and spectral efficiency (32 bps/Hz).

eight $2 \times 2$ stacked Alamouti codes (rate-1) with 16-QAM and IC algorithm in [88],

v) $16 \times 16$ V-BLAST scheme (rate-16) with 4-QAM and sphere decoding (SD) algorithm in [114],

vi) $16 \times 16$ V-BLAST scheme (rate-16) with 4-QAM and ZF-SIC detector, and

vii) $16 \times 16$ V-BLAST scheme (rate-16) with 4-QAM and ISIC algorithm in [111].

We present the BER performance comparison of these different combinations in Fig. 2.14. We also obtained the complexity numbers (in number of real operations per bit) from simulations for these different combinations at an uncoded BER of $5 \times 10^{-2}$; these numbers are presented in Table 2.1, along with the SNRs at which $5 \times 10^{-2}$ uncoded BER is achieved. The following interesting observations can be made from Fig. 2.14 and Table 2.1:

- the proposed scheme (combination i)) significantly outperforms the stacked architecture/IC detector combinations presented in [88] (combinations iii) and iv));

e.g., at $5 \times 10^{-2}$ uncoded BER, the proposed scheme performs better than the stacked architecture/IC in [88] by 17 dB (for four $4 \times 4$ QOSTBCs) and 10 dB (for eight $2 \times 2$ Alamouti codes). Also, the proposed scheme achieves this significant
### Table 2.1: Complexity and performance comparison of different large-MIMO architecture/detector combinations, all with $N_t = N_r = 16$ and achieving 32 bps/Hz spectral efficiency.

<table>
<thead>
<tr>
<th>No.</th>
<th>Large-MIMO Architecture/Detector Combinations</th>
<th>Complexity (in # real operations per bit at $5 \times 10^{-2}$ uncoded BER)</th>
<th>SNR required to achieve $5 \times 10^{-2}$ uncoded BER</th>
</tr>
</thead>
<tbody>
<tr>
<td>i)</td>
<td>16 × 16 ILL-only CDA STBC (rate-16), 4-QAM and 1-LAS detection [Proposed scheme]</td>
<td>$3.473 \times 10^3$</td>
<td>6.8 dB</td>
</tr>
<tr>
<td>ii)</td>
<td>16 × 16 ILL-only CDA STBC (rate-16), 4-QAM and IC algorithm in [111]</td>
<td>$1.187 \times 10^5$</td>
<td>11.3 dB</td>
</tr>
<tr>
<td>iii)</td>
<td>Four 4 × 4 stacked rate-1 QOSTBCs, 256-QAM and IC algorithm in [88]</td>
<td>$5.54 \times 10^6$</td>
<td>24 dB</td>
</tr>
<tr>
<td>iv)</td>
<td>Eight 2 × 2 stacked rate-1 Alamouti codes, 16-QAM and IC algorithm in [88]</td>
<td>$8.719 \times 10^3$</td>
<td>17 dB</td>
</tr>
<tr>
<td>v)</td>
<td>16 × 16 V-BLAST (rate-16) scheme, 4-QAM and SD algorithm in [114]</td>
<td>$4.66 \times 10^4$</td>
<td>7 dB</td>
</tr>
<tr>
<td>vi)</td>
<td>16 × 16 V-BLAST (rate-16) scheme, 4-QAM and V-BLAST detector (ZF-SIC)</td>
<td>$1.75 \times 10^4$</td>
<td>13 dB</td>
</tr>
<tr>
<td>vii)</td>
<td>16 × 16 V-BLAST (rate-16) scheme, 4-QAM and ISIC algorithm in [111]</td>
<td>$7.883 \times 10^3$</td>
<td>10.6 dB</td>
</tr>
</tbody>
</table>

performance advantage at a much lesser complexity than those of the stacked architecture/IC combinations (see Table 2.1).

- the proposed scheme performs slightly better than the V-BLAST/sphere decoder combination (combination vi); 6.8 dB in proposed scheme versus 7 dB in V-BLAST with sphere decoding at $5 \times 10^{-2}$ uncoded BER. Importantly, the proposed scheme enjoys a significant complexity advantage (by more than an order) over the V-BLAST/sphere decoder combination.

- the ISIC algorithm in [111] applied to ILL-only STBC detection (combination ii) is inferior to the proposed scheme in both performance (by about 4.5 dB at $5 \times 10^{-2}$ uncoded BER) as well as complexity (by about two orders).

- the ISIC algorithm in [111] applied to $16 \times 16$ V-BLAST detection (combination
vi(i) is also inferior to the proposed scheme in BER performance (by about 3.8 dB at $5 \times 10^{-2}$ uncoded BER) as well as complexity (by about a factor of 2).

- comparing the stacked architecture/IC combinations with V-BLAST/ZF-SIC (combination vi) and V-BLAST/ISIC combinations, we see that although the diversity orders achieved in stacked architecture/IC combinations are high (see their slopes at high SNRs in Fig. 2.14), V-BLAST with ZF-SIC and ISIC detectors perform much better at low and medium SNRs.

In summary, the proposed scheme outperforms the other considered architecture/detector combinations both in terms of performance as well as complexity.

### 2.4.5 Turbo Coded BER and Nearness-to-Capacity Results

**Turbo Coded BER performance: Fast Fading Channel**

We discuss the turbo coded BER performance of the proposed scheme when the MIMO channel is assumed to be quasi-static (i.e., for a $(n, p, k)$ non-orthogonal STBC, the MIMO channel is static for $p$ channel uses, and changes to an independent realization for the next $p$ channel uses). In all the coded BER simulations, we fed the soft outputs presented in Sec. 2.2.5 as input to the turbo decoder. In Fig. 2.15, we plot the turbo coded BER of the 1-LAS detector in decoding $16 \times 16$ FD-ILL and ILL-only STBCs, with $N_t = N_r = 16$, 4-QAM and turbo code rates $1/3$ (10.6 bps/Hz), $1/2$ (16 bps/Hz), $3/4$ (24 bps/Hz). The minimum SNRs required to achieve these capacities in a $16 \times 16$ MIMO channel (obtained by evaluating the ergodic MIMO capacity expression in [5] through simulation) are also shown. It can be seen that the 1-LAS detector performs close to within just about 4 dB from capacity, which is very good in terms of nearness-to-capacity considering the high spectral efficiencies achieved. It can also be seen that the coded BER performance of FD-ILL and ILL-only STBCs are almost the same for the system parameters considered. Figure 2.16 shows the turbo coded BER
Chapter 2. Large-MIMO Detection Using Likelihood Ascent Search

Figure 2.15: Turbo coded BER performance of 1-LAS detector for $16 \times 16$ FD-ILL and ILL-only STBCs. $N_t = N_r = 16$, 4-QAM, turbo code rates: $1/3$, $1/2$, $3/4$ ($10.6$, $16$, $24$ bps/Hz).

performance of $32 \times 32$ ILL-only STBC with $N_t = N_r = 32$, 16-QAM, and turbo code rates $1/3$ ($42.6$ bps/Hz), $1/2$ ($64$ bps/Hz), $3/4$ ($96$ bps/Hz) using 1-LAS detection.

**Turbo Coded BER Performance: Slow Fading Channel**

We next discuss the turbo coded performance of the proposed 1-LAS detector in a slow fading channel, where the MIMO channel is static for the full duration of the turbo coded frame, and changes to an independent realization in the next turbo frame. Figure 2.17 shows the rate-$3/4$ turbo coded performance of 1-LAS detection with MMSE initial filter for $N_r = N_t = (4, 8, 12)$ and 4-QAM. FD-ILL STBCs are used in order to achieve the full diversity gain of the MIMO channel. The target spectral efficiency is $1.5N_t$ bps/Hz. The MIMO channel outage probability (given by (1.3)), which is the theoretical limit for the codeword error probability (CEP) of long codewords, is also plotted for the sake of comparison. It is observed that, 1-LAS search improves the codeword error probability significantly when compared to the error performance of the MMSE-only initial detector. At a CEP of $10^{-2}$, MMSE 1-LAS performs better than the MMSE-only detector by about 2.5 dB for $N_r = N_t = 8, 12$. Further, with increasing
Figure 2.16: Turbo coded BER performance of 1-LAS detector for $32 \times 32$ ILL-only STBC. $N_t = N_r = 32$, 16-QAM, turbo code rates: 1/3, 1/2, 3/4 (42.6, 64, 96 bps/Hz).

Figure 2.17: Turbo codeword error probability of 1-LAS detection in a slow fading MIMO channel for $N_t = N_r = 4, 8, 12$, 4-QAM, rate-3/4 turbo code and FD-ILL STBCs. MMSE initial vector.
$N_r = N_t$, the error performance of MMSE 1-LAS is observed to improve and achieve a higher diversity slope. This is similar to the ‘large-system behavior’ observed for the uncoded BER performance of the LAS detector. The MMSE 1-LAS detector is also observed to perform close to the theoretical outage probability of the MIMO channel. At a CEP of $10^{-2}$, the MMSE 1-LAS detector performs within 4.5 dB of the MIMO channel outage probability for $N_r = N_t = 12$.

### 2.4.6 Effect of MIMO Spatial Correlation

In generating the BER results in Figs. 2.9 to 2.16, we have assumed i.i.d. fading. However, MIMO propagation conditions witnessed in practice often render the i.i.d. fading model as inadequate. More realistic MIMO channel models that take into account the scattering environment, spatial correlation, etc., have been investigated in the literature [8],[9]. For example, spatial correlation at the transmit and/or receive side can affect the rank structure of the MIMO channel resulting in degraded MIMO capacity [9]. The structure of scattering in the propagation environment can also affect the capacity [8]. Hence, it is of interest to investigate the performance of the $M$-LAS detector in more realistic MIMO channel models. To this end, we use the non-line-of-sight (NLOS) correlated MIMO channel model proposed by Gesbert et al in [8], and evaluate the effect of spatial correlation on the BER performance of the $M$-LAS detector. We note that this model can be appropriate in application scenarios like high data rate wireless IPTV/HDTV distribution using high spectral efficiency large-MIMO links, where large $N_t$ and $N_r$ can be placed at the base station (BS) and customer premises equipment (CPE), respectively. The propagation scenario for the MIMO channel model considered is shown in Fig. 2.18, where linear arrays of $N_t$ omnidirectional transmit antennas with spacing $d_t$, and $N_r$ omnidirectional receive antennas with spacing $d_r$ are considered [8]. The propagation path between the transmit and receive arrays is obstructed on both sides of the link by a number of significant near-field scatterers (e.g., large...
objects) referred to as transmit and receive scatterers, which are modeled as omnidirectional ideal scatterers. The maximum range of the scatterers from the horizontal axis at the transmit and receive sides are denoted by $D_t$ and $D_r$, respectively. When omnidirectional antennas are used, $D_t$ and $D_r$ correspond to the transmit and receive scattering radii, respectively. On the receive side, the signal reflected by the scatterers onto the antennas impinge on the array with an angular spread $\theta_r$, which is a function of the distance between the array and the scatterers. Similarly, angular spread $\theta_t$ is defined on the transmit side. The range between the local scatterers at the transmit and receive sides is denoted by $R$. It is assumed that the scatterers are located adequately far from the antennas so that the plane-wave assumption holds. Further, local scattering condition is assumed, i.e., $D_t << R$ and $D_r << R$. The number of scatterers on each side, $S$, is considered to be large enough (typically $> 10$) for random fading to occur. The complex channel gain matrix as per this model is given by

$$H_c = \frac{1}{\sqrt{S}} R_{\theta_t,d_t}^{1/2} G_t R_{\theta_r,d_r}^{1/2} G_r R_{\theta_t,d_t}^{1/2} G_t R_{\theta_r,d_r}^{1/2} G_r,$$

(2.48)

where $G_t = [g_1 g_2 \cdots g_{N_t}]$ is an $S \times N_t$ i.i.d. Rayleigh fading matrix, $g_m \sim \mathcal{CN}(0, I_S)$, $G_r = [g_1 g_2 \cdots g_{N_r}]$, $R_{\theta_t,d_t}^{1/2}$ and $R_{\theta_r,d_r}^{1/2}$ are the $N_t \times N_t$ and $N_r \times N_r$ matrices controlling the transmit and receive antenna correlations, respectively, whose expressions are given in [8].

We consider the following parameters\footnote{The parameters used in the model in [8] include: $N_t, N_r$: # transmit and receive (omni-directional) antennas; $d_t, d_r$: spacing between antenna elements at the transmit side and at the receive side; $R$: distance between transmitter and receiver; $D_t, D_r$: transmit and receive scattering radii; $S$: number of scatterers on each side; $\theta_t, \theta_r$: angular spread at the transmit and receiver sides, and $f_c, \lambda$: carrier frequency, wavelength.} in the simulations: $f_c = 5$ GHz, $R = 500$ m, $S = 30$, $D_t = D_r = 20$ m, $\theta_t = \theta_r = 90^\circ$, and $d_t = d_r = 2\lambda/3$. For $f_c = 5$ GHz, $\lambda = 6$ cm and $d_t = d_r = 4$ cm. In Fig. 2.14, we plot the BER performance of the 1-LAS detector in decoding $16 \times 16$ ILL-only STBC with $N_t = N_r = 16$ and 16-QAM. Uncoded BER as well as rate-3/4 turbo coded BER (48 bps/Hz spectral efficiency) for i.i.d. fading.
as well as correlated fading are shown. In addition, from the MIMO capacity formula in [5], we evaluated the theoretical minimum SNRs required to achieve a capacity of 48 bps/Hz in i.i.d. as well as correlated fading, and plotted them also in Fig. 2.14. It is seen that the minimum SNR required to achieve a certain capacity (48 bps/Hz) gets increased for correlated fading compared to i.i.d. fading. From the BER plots in Fig. 2.14, it can be observed that at an uncoded BER of $10^{-3}$, the performance in correlated fading degrades by about 7 dB compared that in i.i.d. fading. Likewise, at a rate-3/4 turbo coded BER of $10^{-4}$, a performance loss of about 6 dB is observed in correlated fading compared to that in i.i.d. fading. In terms of nearness to capacity, the vertical fall of the coded BER for i.i.d. fading occurs at about 24 dB SNR, which is about 13 dB away from theoretical minimum required SNR of 11.1 dB. With correlated fading, the detector is observed to perform close to capacity within about 18.5 dB. One way to alleviate such degradation in performance due to spatial correlation can be by providing more number of dimensions at the receive side, which is highlighted in Fig. 2.19.

Figure 2.19 illustrates that the 1-LAS detector can achieve substantial improvement in uncoded as well as coded BER performance in decoding $12 \times 12$ ILL-only STBC by increasing $N_r$ beyond $N_t$ for 16-QAM in correlated fading. In the simulations,
we have maintained $N_r d_r = 72 \text{ cm}$ and $d_t = d_r$ in both the cases of symmetry (i.e., $N_t = N_r = 12$) as well as asymmetry (i.e., $N_t = 12$, $N_r = 18$). By comparing the 1-LAS detector performance with $[N_t = N_r = 12]$ versus $[N_t = 12$, $N_r = 18]$, we observe that the uncoded BER performance with $[N_t = 12$, $N_r = 18]$ improves by about 17 dB compared to that of $[N_t = N_r = 12]$ at $2 \times 10^{-3}$ BER. Even the uncoded BER performance with $[N_t = 12$, $N_r = 18]$ is significantly better than the coded BER performance with $[N_t = N_r = 12]$ by about 11.5 dB at $10^{-3}$ BER. This improvement is essentially due to the ability of the 1-LAS detector to effectively pick up the additional diversity orders provided by the increased number of receive antennas. With a rate-3/4 turbo code (i.e., 36 bps/Hz), at a coded BER of $10^{-4}$, the 1-LAS detector achieves a significant performance improvement of about 13 dB with $[N_t = 12$, $N_r = 18]$ compared to that with $[N_t = N_r = 12]$. With $[N_t = 12$, $N_r = 18]$, the vertical fall of coded BER is such that it is only about 8 dB from the theoretical minimum SNR needed to achieve capacity. This points to the potential for realizing high spectral efficiency multi-gigabit large-MIMO
systems that can achieve good performance even in the presence of spatial correlation. We further remark that transmit correlation in MIMO fading can be exploited by using non-isotropic inputs (precoding) based on the knowledge of the channel correlation matrices [10]-[12]. While [10]-[12] propose precoders in conjunction with orthogonal/quasi-orthogonal small MIMO systems in correlated Rayleigh/Ricean fading, design of precoders for large-MIMO systems can be investigated as future work.

2.5 Iterative Detection/Channel Estimation

In this section, we relax the perfect CSIR assumption made in the previous section, and estimate the channel matrix based on a training-based iterative detection/channel estimation scheme. Training-based schemes, where a pilot signal known to the transmitter and the receiver is sent to get a rough estimate of the channel (training phase) has been studied for STBC MIMO systems in [115]-[118]. Here, we adopt a training-based approach for channel estimation in large STBC MIMO systems. In the considered training-based channel estimation scheme, transmission is carried out in frames, where one $N_t \times N_t$ pilot matrix, $X^{(P)}_c \in \mathbb{C}^{N_t \times N_t}$, for training purposes, followed by $N_d$ data STBC matrices, $X^{(i)}_c \in \mathbb{C}^{N_t \times N_t}$, $i = 1, 2, ..., N_d$, are sent in each frame as shown in Fig. 2.20. One frame length, $T$, (taken to be the channel coherence time) is $T = (N_d + 1)N_t$ channel uses. A frame of transmitted pilot and data matrices is of dimension $N_t \times N_t(1 + N_d)$, which can be written as

$$X^c = \begin{bmatrix} X^{(P)}_c & X^{(1)}_c & X^{(2)}_c & \cdots & X^{(N_d)}_c \end{bmatrix}. \quad (2.49)$$

As in [119], let $\gamma_p$ and $\gamma_d$ denote the average SNR during pilot and data phases, respectively, which are related to the average received SNR $\gamma$ as $\gamma(N_d + 1) = \gamma_p + N_d \gamma_d$. Define $\beta_p \triangleq \frac{\gamma_p}{\gamma}$ and $\beta_d \triangleq \frac{\gamma_d}{\gamma}$. Let $E_s$ denote the average energy of the transmitted symbol during the data phase. The average received signal power during the data
phase is given by $\mathbb{E}[\text{tr}(X_i^{(j)}X_i^{(j)H})] = N_t^2 E_s$, and the average received signal power during the pilot phase is $\mathbb{E}[\text{tr}(X_c^{(p)}X_c^{(p)H})] = \frac{N_t^2 E_s \beta_p}{\beta_d} = \mu N_t$, where $\mu \equiv \frac{N_t E_s \beta_p}{\beta_d}$. For optimal training, the pilot matrix should be such that $X_c^{(p)}X_c^{(p)H} = \mu I_{N_t}$ [119]. As in Section 2.1.2, let $H_c \in \mathbb{C}^{N_r \times N_t}$ denote the channel matrix, which we want to estimate. We assume block fading, where the channel gains remain constant over one frame consisting of $(1 + N_d)N_t$ channel uses, which can be viewed as the channel coherence time. This assumption can be valid in slow fading fixed wireless applications (e.g., as in possible applications like BS-to-BS backbone connectivity and BS-to-CPE wireless IPTV/HDTV distribution). For this training-based system and channel model, Hassibi and Hochwald presented a lower bound on the capacity in [119]; we will illustrate the nearness of the performance achieved by the proposed iterative detection/estimation scheme to this bound. The received frame is of dimension $N_r \times N_c(1 + N_d)$, and can be written as

$$Y_c = [Y_c^{(P)} Y_c^{(1)} Y_c^{(2)} \cdots Y_c^{(N_d)}] = H_c X_c + N_c,$$  \hspace{1cm} (2.50)

where $N_c = [N_c^{(P)} N_c^{(1)} N_c^{(2)} \cdots N_c^{(N_d)}]$ is the $N_r \times N_t(1 + N_d)$ noise matrix and its entries are modeled as i.i.d. $\mathcal{C}\mathcal{N}(0, \sigma^2 = \frac{N_t E_s}{\gamma \beta_d})$. Equation (2.50) can be decomposed into two parts, namely, the pilot matrix part and the data matrices part, as

$$Y_c^{(P)} = H_c X_c^{(P)} + N_c^{(P)};$$  \hspace{1cm} (2.51)

$$Y_c^{(D)} = [Y_c^{(1)} Y_c^{(2)} \cdots Y_c^{(N_d)}]$$

$$= H_c [X_c^{(1)} X_c^{(2)} \cdots X_c^{(N_d)}] + [N_c^{(1)} N_c^{(2)} \cdots N_c^{(N_d)}].$$  \hspace{1cm} (2.52)
Figure 2.20: Transmission scheme with one pilot matrix followed by $N_d$ data STBC matrices in each frame.

### 2.5.1 MMSE Estimation Scheme

A straightforward way to achieve detection of data symbols with estimated channel coefficients is as follows:

1. Estimate the channel gains via an MMSE estimator from the signal received during the first $N_t$ channel uses (i.e., during pilot transmission); i.e., given $Y_c^{(P)}$ and $X_c^{(P)}$, an estimate of the channel matrix $H_c$ is found as

$$H_{c^{est}} = Y_c^{(P)} (X_c^{(P)})^H \left[ \sigma^2 I_{N_t} + X_c^{(P)} (X_c^{(P)})^H \right]^{-1}.$$  \hspace{1cm} (2.53)

2. Use the above $H_{c^{est}}$ in place of $H_c$ in the LAS algorithm (as described in Section 2.2) and detect the transmitted data symbols.

We refer to the above scheme as the ‘MMSE estimation scheme.’ In the absence of the knowledge of $\sigma^2$, a zero-forcing estimate can be obtained at the cost of some performance loss compared to the MMSE estimate. The performance of the estimator can be improved by using a cyclic minimization technique for minimizing the ML metric [120].

### 2.5.2 Proposed Iterative Detection/Estimation Scheme

Techniques that employ iterations between channel estimation and detection can offer improved performance. Iterative receiver algorithms are attractive to achieve a good
tradeoff between performance and complexity [121]-[125]. In [121]-[123], receivers that iterate between channel estimation, multiuser detection and channel decoding in coded CDMA systems are presented. Similar iterative techniques in the context of MIMO and MIMO-OFDM systems are presented in [15]-[125]. Here, we propose an iterative scheme, where we iterate between channel estimation and detection in the considered large STBC MIMO system. The proposed scheme works as follows:

1. Obtain an initial estimate of the channel matrix using the MMSE estimator in (2.53) from the pilot part.

2. Using the estimated channel matrix, detect the data STBC matrices $X_c^{(i)}$, $i = 1, 2, \cdots, N_d$ using the LAS detector. Substituting these detected STBC matrices into (2.49), form $X_c^{est}$.

3. Re-estimate the channel matrix using $X_c^{est}$ from the previous step, via

   \[
   H_c^{est} = Y_c(X_c^{est})^H \left[ \sigma^2 I_{N_t} + X_c^{est}(X_c^{est})^H \right]^{-1}.
   \]  

4. Iterate steps 2 and 3 for a specified number of iterations.

The total complexity of obtaining the MMSE estimate of the channel matrix $H_c^{est}$ in (2.53) and (2.54) is $O(N_t^2 N_r) + O(N_t^3)$, which is less than the total complexity of 1-LAS detection of $O(N_t^4 \log N_t)$ for ILL-only STBCs.

### 2.5.3 BER Performance with Estimated CSIR

We evaluated the BER performance of the 1-LAS detector using estimated CSIR, where we estimate the channel gain matrix through the training-based estimation schemes described in the previous two subsections. We consider the BER performance under three scenarios, namely, (i) under perfect CSIR, (ii) under CSIR estimated using the MMSE estimation scheme in Section 2.5.1, and (iii) under CSIR estimated using the
iterative detection/estimation scheme in Section 2.5.2. In the case of estimated CSIR, we show plots for $1P+N_d D$ training, where by $1P+N_d D$ training we mean a training scheme with a frame size of $1 + N_d$ matrices, with 1 pilot matrix followed $N_d$ data STBC matrices from CDA. For this $1P+N_d D$ training scheme, a lower bound on the capacity is given by [119]

$$ C \geq \frac{T - \tau}{T} \mathbb{E} \left[ \log \text{det} \left( I_{N_t} + \frac{\gamma^2 \beta_d \beta_p \tau}{N_t (1 + \gamma \beta_d) + \gamma \beta_p \tau} \hat{H}_c \hat{H}_c^H \right) \right], \quad (2.55) $$

where $T$ and $\tau$, respectively, are the frame size (i.e., channel coherence time) and pilot duration in number of channel uses, and $\sigma^2_{\hat{H}_c} = \frac{1}{N_t N_r} \mathbb{E} [\text{tr} \{ \hat{H}_c \hat{H}_c^H \}]$, where $\hat{H}_c = \mathbb{E}[\hat{H}_c|X_c^{(p)}, Y_c^{(p)}]$ is the MMSE estimate of the channel gain matrix. We computed the capacity bound in (2.55) through simulations for $1P+8D$ and $1P+1D$ training for a $16 \times 16$ MIMO channel. For $1P+8D$ training $T = (1 + 8)16 = 144, \tau = 16$, and for $1P+1D$ training $T = (1 + 1)16 = 32, \tau = 16$. In computing the bounds (shown in Fig. 2.21) and in BER simulations (in Figs. 2.22 and 2.23), we have used $\beta_p = \beta_d = 1$. In Fig. 2.21, we plot the computed capacity bounds, along with the capacity under perfect CSIR.

Figure 2.21: Hassibi-Hochwald (H-H) capacity bound for $1P+8D$ ($T = 144, \tau = 16, \beta_p = \beta_d = 1$) and $1P+1D$ ($T = 32, \tau = 16, \beta_p = \beta_d = 1$) training for a $16 \times 16$ MIMO channel. Perfect CSIR capacity is also shown.
Figure 2.22: Uncoded BER of 1-LAS detector for 16 × 16 ILL-only STBC with i) perfect CSIR, ii) CSIR using MMSE estimation scheme, and iii) CSIR using iterative detection/channel estimation scheme (4 iterations). $N_t = N_r = 16$, 4-QAM, 1P+1D ($T = 32, \tau = 16, \beta_p = \beta_d = 1$) and 1P+8D ($T = 144, \tau = 16, \beta_p = \beta_d = 1$) training.

We obtain the minimum SNR for a given capacity bound in (2.55) from the plots in Fig. 2.21, and show (later in Fig. 2.23) the nearness of the coded BER of the proposed scheme to this SNR limit. We note that improved capacity and BER performance can be achieved if optimum pilot/data power allocation derived in [119] is used instead of the allocation used in Figs. 2.21 to 2.23 (i.e., $\beta_p = \beta_d = 1$). We have used the optimum power allocation in [119] for generating the BER plots in Figs. 2.24 and 2.25. In all the BER simulations with training, $\sqrt{\mu} I_{N_t}$ is used as the pilot matrix. ILL-only STBCs and 1-LAS detection are used.

First, in Fig. 2.22, we plot the uncoded BER performance of 1-LAS detector when 1P+1D and 1P+8D training are used for channel estimation in a 16 × 16 STBC MIMO system with $N_t = N_r = 16$ and 4-QAM. BER performance with perfect CSIR is also plotted for comparison. From Fig. 2.22, it can be observed that, as expected, the BER degrades with estimated CSIR compared to that with perfect CSIR. With MMSE estimation scheme, the performance with 1P+1D and 1P+8D are same because of the
Figure 2.23: Turbo coded BER performance of 1-LAS detector for 16×16 ILL-only STBC with i) perfect CSIR, ii) CSIR using MMSE estimation, and iii) CSIR using iterative detection/channel estimation (4 iterations). \( N_t = N_r = 16 \), 4-QAM, rate-3/4 turbo code, 1P+1D \( (T = 32, \tau = 16, \beta_p = \beta_d = 1) \) and 1P+8D \( (T = 144, \tau = 16, \beta_p = \beta_d = 1) \) training.

one-shot estimation. Also, with 1P+1D training, both the MMSE estimation scheme as well as the iterative detection/estimation scheme (with 4 iterations between detection and estimation) perform almost the same, which is about 3 dB worse compared to that of perfect CSIR at an uncoded BER of \( 10^{-3} \). This indicates that with 1P+NdD training, iteration between detection and estimation does not improve performance much over the non-iterative scheme (i.e., the MMSE estimation scheme) for small \( Nd \). With large \( Nd \) (e.g., slow fading), however, the iterative scheme outperforms the non-iterative scheme; e.g., with 1P+8D training, the performance of the iterative detection/estimation improves by about 1 dB compared to the MMSE estimation.

Next, in Fig. 2.23, we present the rate-3/4 turbo coded BER of 1-LAS detector using estimated CSIR for the cases of 1P+8D and 1P+1D training. From Fig. 2.23, it can be seen that, compared to that of perfect CSIR, the estimated CSIR performance is worse by about 3 dB in terms of coded BER for 1P+8D training. With MMSE estimation scheme,
Figure 2.24: Turbo coded BER performance of 1-LAS detection and iterative estimation/detection as a function of coherence time, $T = 32, 144, 400, 784$, for a given $N_t = N_r = 16$, $16 \times 16$ ILL-only STBC, 4-QAM, rate-3/4 turbo code. $10^{-4}$ coded BER occurs at about $12 - 7.7 = 4.3$ dB away from the capacity bound for 1P+1D and 1P+8D training. This nearness to capacity bound improves by about 0.6 dB for the iterative detection/estimation scheme. We note that for the system in Fig. 2.23 with parameters $16 \times 16$ STBC, 4-QAM, rate-3/4 turbo code, and 1P+8D training with $T = 144, \tau = 16$, we achieve a high spectral efficiency of $16 \times 2 \times \frac{3}{4} \times \frac{8}{5} = 21.3$ bps/Hz even after accounting for the overheads involved in channel estimation (i.e., pilot matrix) and channel coding, while achieving good near-capacity performance at low complexity. This points to the suitability of the proposed approach of using LAS detection along with iterative detection/estimation in practical implementation of large STBC MIMO systems.

Finally, in Fig. 2.24, we illustrate the coded BER performance of 1-LAS detection and iterative detection/estimation scheme for different coherence times, $T$, for a fixed $N_t = N_r = 16$, $16 \times 16$ STBC, 4-QAM, and rate-3/4 turbo code. The various values of $T$ considered and the corresponding spectral efficiencies are: $i$) $T = 32$, 1P+1D, 12 bps/Hz, $ii$) $T = 144$, 1P+8D, 21.3 bps/Hz, $iii$) $T = 400$, 1P+24D, 23.1 bps/Hz, and
**iv)** $T = 784$, 1P+48D, 23.5 bps/Hz. In all these cases, the corresponding optimum pilot/data power allocations in [119] are used. From Fig. 2.24, it can be seen that for these four cases, $10^{-4}$ coded BER occurs at around 12 dB, 10.6 dB, 9.7 dB, and 9.4 dB, respectively. The $10^{-4}$ coded BER for perfect CSIR happens at around 8.5 dB. This indicates that the performance with estimated CSIR improves as $T$ is increased, and that a performance loss of less than 1 dB compared to perfect CSIR can be achieved with large $T$ (i.e., slow fading). For example, with 1P+48D training ($T = 784$), the performance with estimated CSIR gets close to that with perfect CSIR both in terms of spectral efficiency (23.5 *vs* 24 bps/Hz) as well as SNR at which $10^{-4}$ coded BER occurs (8.5 *vs* 9.4 dB). This is expected, since the channel estimation becomes increasingly accurate in slow fading (large coherent times) while incurring only a small loss in spectral efficiency due to pilot matrix overhead. This result is significant because $T$ is typically large in fixed/low-mobility wireless applications, and the proposed system can effectively achieve high spectral efficiencies as well as good performance in such applications.
Table 2.2: On optimum $N_t$ for a given $N_r$ and $T$. System-II with a smaller $N_t$ achieves a higher spectral efficiency while achieving $10^{-3}$ coded BER at a lesser SNR than System-I with a larger $N_t$.

### 2.5.4 On Optimum $N_t$ for a Given $N_r$ and $T$

In [119], through theoretical capacity bounds it has been shown that, for a given $N_r$, $T$ and SNR, there is an optimum value of $N_t$ that maximizes the capacity bound (refer Figs. 5 and 6 in [119], where the optimum $N_t$ is shown to be greater than $N_r$ in Fig. 5 and less than $N_r$ in Fig. 6). For example, for $N_r = 16$, $T = 48$, and SNR = 10 dB, the capacity bound evaluated using (2.55) with optimum power allocation for $N_t = 12$ is 19.73 bps/Hz, whereas for $N_t = 16$ the capacity bound reduces to 17.53 bps/Hz showing that the optimum $N_t$ in this case will be less than $N_r$. We demonstrate such an observation in practical systems by comparing the simulated coded BER performance of two systems, referred to as System-I and System-II, using 1-LAS detection and iterative detection/estimation scheme. The parameters of System-I and System-II are listed in Table 2.2. $N_r$ and $T$ are fixed at 16 and 48, respectively, in both systems. System-I uses 16 transmit antennas and $16 \times 16$ STBC, whereas System-II uses 12 transmit antennas and $12 \times 12$ STBC. Since the pilot matrix is $\sqrt{\mu} I_{N_r}$, the pilot duration $\tau$ is 16 and 12, respectively, for System-I and System-II. Optimum pilot/data power allocation and
4-QAM modulation are employed in both systems. System-I uses rate-1/2 turbo code and system-II uses rate-3/4 turbo code. With the above system parameters, the spectral efficiency achieved in System-I is \(16 \times \frac{1}{2} \times \frac{2}{3} = 10.33\) bps/Hz, whereas System-II achieves a higher spectral efficiency of \(12 \times 2 \times \frac{3}{4} \times \frac{3}{4} = 13.5\) bps/Hz. In Fig. 2.25, we plot the coded BER of both these systems using 1-LAS detection and iterative detection/estimation. From the simulation points shown in Fig. 2.25, it can be observed that System-II with a smaller \(N_t\) and higher spectral efficiency in fact achieves a certain coded BER performance at a lesser SNR compared to System-I. For example, to achieve \(10^{-3}\) coded BER, System-I requires an SNR of about 8.9 dB, whereas System-II requires only 8.6 dB. This implies that because of the reduction of throughput due to pilot symbols (by a factor of \(\frac{T-\tau}{T}\) for a given \(T\) and \(\tau = N_t\)), a larger \(N_t\) does not necessarily mean a higher spectral efficiency. Such an observation has also been made in [119] based on theoretical capacity bounds. The proposed detection/channel estimation scheme allows the prediction of such behavior through simulations, which, in turn, allows system designers to find optimum \(N_t\) and STBC size to achieve a certain spectral efficiency in large STBC MIMO systems.
Chapter 3

Large-System Performance Analysis of LAS Algorithm

In Chapter 2, we presented the LAS detection algorithm, and showed through simulations that its error performance improves with increasing $N_t = N_r$. The BER performance indeed was seen to get increasingly closer to SISO AWGN performance for increasingly large $N_t = N_r$ at moderate to high SNRs, indicating optimality in the large-system limit. In this chapter, in an attempt to gain some insights on the large-system behavior of the algorithm, we carry out an asymptotic performance analysis of the LAS algorithm in the large-system limit, where $N_t \to \infty$ keeping $N_t = N_r$. We present the analysis for a V-BLAST MIMO system, whose real-valued system model (as described in Chapter 2) is given by

$$y = Hx + n, \quad (3.1)$$

where $H \in \mathbb{R}^{2N_r \times 2N_t}$ is the equivalent channel matrix, $y \in \mathbb{R}^{2N_r}$ and $x \in \mathbb{R}^{2N_t}$ are the received and transmitted vectors respectively, and $n \in \mathbb{R}^{2N_r \times 1}$ is the additive Gaussian noise vector. We consider 4-QAM modulation. So, the set of possible input symbols is $S = \{+1, -1\}^{2N_r}$. For a given received vector $y$ and channel realization $H$, the output
of the ML detector is given by
\[
d_{ML} = \arg \min_{d \in S} \| y - Hd \|^2 = \arg \min_{d \in S} \left( d^T H^T H d - 2 y^T H d \right).
\] (3.2)

### 3.1 Asymptotic Analysis of the LAS Algorithm

In this section, we analyze the error performance of the LAS algorithm in V-BLAST MIMO for \( N_t, N_r \to \infty \), with \( N_t = N_r \) and 4-QAM, i.e., \( S = \{+1, -1\}^{2N_t} \).

An \( n \)-symbol update on a data vector \( d \in S \) transforms \( d \) to \( (d - \Delta d_n) \) such that \( (d - \Delta d_n) \in S \). Further, \( (d - \Delta d_n) \) is obtained by changing \( n \) symbols in \( d \) at distinct indices given by the \( n \)-tuple \( u_n \triangleq (i_1, i_2, \cdots, i_n), 1 \leq i_j \leq 2N_t, \forall j = 1, \cdots, n \) and \( i_j \neq i_k \) for \( j \neq k \). Therefore, we can write \( \Delta d_n \) as
\[
\Delta d_n = \sum_{k=1}^{n} 2d_{i_k} e_{i_k},
\] (3.3)

where \( d_{i_k} \) is the \( i_k \)-th element of \( d \) and \( e_{i_k} \) is a vector with its \( i_k \)-th entry only as one and all other entries as zero. Let \( \mathbb{L}_n \subseteq S \) denote the set of data vectors such that for any \( d \in \mathbb{L}_n \), if a \( n \)-symbol update is performed on \( d \) resulting in a vector \( (d - \Delta d_n) \), then
\[
\| y - H(d - \Delta d_n) \| \geq \| y - H d \|.
\]

**Lemma 1.** Let \( d \in S \). Then, \( d \in \mathbb{L}_n \) if and only if, for any \( n \)-update on \( d \), \( n \in [1, 2, \cdots, 2N_t] \),
\[
(y - H d + \frac{1}{2} H \Delta d_n)^T (H \Delta d_n) \geq 0.
\] (3.4)

**Proof:** By definition, if \( d \in \mathbb{L}_n \), then no \( n \)-symbol update can result in a reduction in the ML cost function. Using this, we can write
\[
\| y - H(d - \Delta d_n) \|^2 \geq \| y - H d \|^2.
\] (3.5)

Simplifying (3.5), we get (3.4). Since the choice of the indices in \( u_n \) is arbitrary, the
Chapter 3. Large-System Performance Analysis of LAS Algorithm

Lemma holds true for all possible \( n \)-tuples of distinct indices. For the converse, if \( d \) satisfies (3.4) for all possible \( u_n \) for a given \( n \), then, since (3.4) and (3.5) are equivalent, \( d \) also satisfies (3.5) for all possible \( u_n \). This implies that \( d \in \mathbb{L}_n \). □

If \( d \in \mathbb{L}_1 \), then using Lemma 1 and (3.1), we can write

\[
(n + H(x - d) + h_p d_p)^T h_p d_p \geq 0, \quad \forall p = 1, \ldots, 2N_t,
\]

where \( h_p \) is the \( p \)-th column of \( H \).

**Lemma 2.** Assuming uniqueness of the ML vector \( d_{ML} \) in (3.2), a symbol vector \( d \in S \) is the ML vector if and only if the noise vector \( n \) satisfies the following set of equations

\[
\left( n + H(x - d) + \left( \sum_{j=1}^{n} h_{ij} d_{ij} \right) \right)^T \left( \sum_{j=1}^{n} h_{ij} d_{ij} \right) \geq 0, \quad \forall \ n = 1, \ldots, 2N_t,
\]

for all possible \( n \)-tuples \((i_1, \ldots, i_n)\) for each \( n \).

**Proof:** If \( d \) is the unique ML vector, then from the definition of the ML criterion in (3.2), it must be true that any \( n \)-update on \( d \) will not result in any decrease in the ML cost function. Therefore, \( d \in \mathbb{L}_n, \forall \ n = 1, 2 \cdots, 2N_t \). Hence, by Lemma 1, it must be true that \( d \) satisfies (3.4) for all \( n = 1, 2, \cdots, 2N_t \) and for all possible \( u_n \) for each \( n \).

Substituting \( y = Hx + n \) in (3.4), we get (3.7). This proves the direct result. To prove the converse, let the noise vector \( n \) satisfy (3.7) for some vector \( d \). Since \( y = Hx + n \), the conditions in (3.7) imply the conditions in (3.4) for all \( n = 1, 2, \cdots, 2N_t \) and for all possible \( u_n \) for each \( n \). Therefore, by Lemma 1, \( d \in \mathbb{L}_n \) for all \( n = 1, 2, \cdots, 2N_t \), which then implies that \( d \) indeed is the ML vector. □

**Definition:** For each \( d \in S \) and for each integer \( m, 1 \leq m \leq 2N_t \), we associate the set of vectors \( \mathcal{R}_{d^m} = \left\{ v \mid v \in \mathbb{R}^{2N_t}, \text{ and } (v + H(x - d) + \left( \sum_{j=1}^{n} h_{ij} d_{ij} \right))^T \left( \sum_{j=1}^{n} h_{ij} d_{ij} \right) \geq 0, \forall n = 1, \cdots, m, \text{ and for all possible } n \text{-tuples } (i_1, \cdots, i_n) \text{ for each } n \right\} \), and define \( \mathcal{R}_d \triangleq \mathcal{R}_{d^{2N_t}} \).

Here we note that, for a given \( x, H \) and \( n \), if \( n \in \mathcal{R}_{d^m} \) for some vector \( d \), then \( d \) is
Lemma 3. If the noise vector \( n \in \mathcal{R}_d \), then \( d \) is the ML vector. Let \( d_i, d_j \in S \) and \( d_i \neq d_j \). Then \( \mathcal{R}_{d_i} \) and \( \mathcal{R}_{d_j} \) are disjoint.

Proof: From Lemma 2 and the definition of \( \mathcal{R}_d \), it is clear that \( d \) is the ML vector if and only if \( n \in \mathcal{R}_d \). The disjointness of \( \mathcal{R}_{d_i} \) and \( \mathcal{R}_{d_j} \), \( i \neq j \), can be shown by contradiction. If \( \mathcal{R}_{d_i} \) and \( \mathcal{R}_{d_j} \) are not disjoint, then there exists some vector \( v \) belonging to both \( \mathcal{R}_{d_i} \) and \( \mathcal{R}_{d_j} \). If \( v \) were to be the noise vector \( n \), then, \( v \) would satisfy the set of equations in (3.7) for both \( d = d_i \) and \( d = d_j \), since \( v \) belongs to both \( \mathcal{R}_{d_i} \) and \( \mathcal{R}_{d_j} \). This, by Lemma 2, implies that both \( d_i \) and \( d_j \) are ML vectors, which is a contradiction because of the uniqueness of the ML vector. □

Lemma 4. Let \( h \in \mathbb{R}^{2N_t} \) be a random vector with i.i.d entries distributed as \( \mathcal{N}(0,0.5) \). Let \( \{h_i\}, i = 1, 2, \ldots, m \) be a set of vectors, with each \( h_i \in \mathbb{R}^{2N_t} \) and having i.i.d entries distributed as \( \mathcal{N}(0,0.5) \), \( \mathbb{E}[h_i h_j^T] = 0 \) for \( i \neq j \), and \( \mathbb{E}[hh_i^T] = 0 \) for \( j = 1, \ldots, m \). Then

\[
\lim_{N_t \to \infty} \frac{\sum_{k=1}^{m} h_i^T h_k}{mN_t} = 0.
\]

(3.8)

Proof: Let \( \tilde{h} \triangleq \frac{1}{\sqrt{m}} \sum_{k=1}^{m} h_k \). Then, \( \tilde{h} \sim \mathcal{N}(0, \frac{1}{2}) \). Therefore, we have

\[
\lim_{N_t \to \infty} \frac{\sum_{k=1}^{m} h_i^T h_k}{mN_t} = \lim_{N_t \to \infty} \frac{h_i^T \tilde{h}}{\sqrt{mN_t}}.
\]

(3.9)

We can write

\[
\lim_{N_t \to \infty} \frac{h_i^T \tilde{h}}{N_t} = \lim_{N_t \to \infty} \frac{\sum_{k=1}^{2N_t} h_k \tilde{h}_k}{N_t},
\]

(3.10)

where \( h_k \) and \( \tilde{h}_k \) are the \( k \)th elements of \( h \) and \( \tilde{h} \), respectively. The r.v’s \( h_k \tilde{h}_k, k = 1, \ldots, 2N_t \) are i.i.d with mean zero. From the strong law of large numbers [126], it follows that \( \lim_{N_t \to \infty} \sum_{k=1}^{2N_t} \frac{h_k \tilde{h}_k}{2N_t} = 0 \). Using this in (3.9) completes the proof. □
Before we present the next lemma, we present the Slutsky’s theorem on convergence of random variables, which is used to prove Lemma 5.

**Slutsky’s Theorem [126]:** Let \( \{X_m\} \) and \( \{Y_m\} \) be sequences of random variables. If \( \{X_m\} \) converges in distribution to a random variable \( X \), and \( \{Y_m\} \) converges in probability to a constant \( c \), then it is true that i) \( \{X_m + Y_m\} \) converges in distribution to \( X + c \), ii) \( \{X_m Y_m\} \) converges in distribution to \( cX \), and iii) \( \{X_m/Y_m\} \) converges in distribution to \( X/c \).

**Lemma 5.** For a given \( u_n \) and a given \( d \in S \), define a r.v \( z_{u_n,d} \) as

\[
 z_{u_n,d} \triangleq \frac{\sum_{k=1}^{n} \sum_{j=k+1}^{n} h_{ij}^T h_{ij} d_i d_j}{\sum_{j=1}^{n} \|h_{ij}\|^2},
\]

(3.11)

where \( i_j \in u_n, j = 1, \cdots, n \). For any \( u_n \) and any \( d \in S \), \( z_{u_n,d} \) converges to zero in probability as \( N_t \to \infty \), i.e., \( z_{u_n,d} \xrightarrow{p} 0 \) as \( N_t \to \infty \), \( \forall n = 2, 3, \cdots, 2N_t \).

**Proof:** Proof of this Lemma is given in Appendix B. □

**Simulated pdf of** \( z_{u_n,d} \): In Fig. 3.1, we plot the simulated pdf of \( z_{u_n,d} \) for \( n = 2N_t \) for different values of \( N_t = N_r \) for a certain \( u_n \) and \( d \) (the pdf was observed to be same for different \( u_n \) and \( d \)). We observe that with increasing \( N_t = N_r \), the pdf of \( z_{u_n,d} \) tends towards the Dirac delta function at zero. This implies that \( z_{u_n,d} \) tends to zero in distribution, and hence in probability, for large \( N_t = N_r \), which was formally proved in Lemma 5.

Next, let \( I(\text{event}) \) denote the indicator function, where \( I(\text{event}) = 1 \) if \( \text{event} \) is true, and \( = 0 \) if \( \text{event} \) is false. For a given \( H, x \) and \( n \) and any positive integer \( 1 \leq m \leq 2N_t \), we define the set \( D_{H,x,n}^m \) as

\[
 D_{H,x,n}^m \triangleq \left\{ d \mid d \in \{-1, +1\}^{2N_t}, n \in \mathcal{R}_{d^m} \right\}.
\]

(3.12)

Note that \( \mathcal{R}_{d^m} \) depends on \( H \) and \( x \). Also, for a given \( H, x \) and \( n \), any vector belonging to \( D_{H,x,n}^m \) is actually an \( m \)-update local minima. It is also obvious that a \( 2N_t \)-update local minima is actually a ML data vector (solution to the optimization problem in (3.2)).
Figure 3.1: Simulated pdf of $z_{u,n,d}$ for $n = 2N_t$ for increasing $N_t = N_r$. 4-QAM. The pdf tends towards Dirac delta function at zero.

We define a formal notion of a detector, which, for a given $H$ and $y$, outputs an estimate of the transmitted vector $x$. Specifically, we consider a family of such detectors, such that for any $N_t \geq 1$ and each $1 \leq m \leq 2N_t$, we have a class of detectors $A_m$, whose output $A_m(H,y)$ is a data vector in $\mathbb{D}^m_{H,x,n}$. Basically, $A_m$ converges to an $m$-update local minima. We shall be interested in detectors $A_1$ which satisfy

$$A_1(H,y) \in \mathbb{D}^1_{H,x,n}. \quad (3.13)$$

The 1-LAS detector presented in Chapter 2, is an example of detector $A_1$. For a given detector $A_1$, let

$$p_m(x,n) \triangleq \mathbb{E}_H[I(n \in \mathcal{R}_{d^m} \mid H,x,n,d = A_1(H,y))]$$

$$\bar{p}_m(x,n) \triangleq \mathbb{E}_H[I(n \notin \mathcal{R}_{d^m} \mid H,x,n,d = A_1(H,y))]$$

$$= 1 - \mathbb{E}_H[I(n \in \mathcal{R}_{d^m} \mid H,x,n,d = A_1(H,y))]. \quad (3.14)$$
For $m = 2N_t$, we simply drop the subscript and define

$$
p(x, n) \triangleq p_{2N_t}(x, n), \quad \bar{p}(x, n) \triangleq \bar{p}_{2N_t}(x, n).
$$

(3.15)

Therefore, for a given $(x, n)$, $p_m(x, n)$ denotes the probability that the output of the detector $A_1$ is an $m$-update local minima. We also define $p_{r,s}$ as

$$
p_{r,s} \triangleq \mathbb{E}(x, n) \mathbb{E}_H[I(n \in \mathcal{R}_{d_s} \mid H, x, n \in \mathcal{R}_{d_r}, d = A_1(H, y))]
$$

(3.16)

i.e., the average probability that the output of the detector $A_1$ is a $s$-update local minima, conditioned on the fact that the output is a $r$-update local minima. From the definition of $\mathcal{R}_{d_m}$, it is obvious that $p_{r,s}$ is 1 for $s \leq r$.

We conjecture that a 1-update local minima is indeed a $2N_t$-update local minima with high probability, which is more formally stated as follows.

**Conjecture 1.** For any detector $A_1$ with property (3.13) and any $\delta, 0 \leq \delta \leq 1$, there exists an integer $N(\delta)$ such that for $N_t > N(\delta)$, and any $(x, n)$, $p(x, n) > 1 - \delta$.

In Appendix C, we present arguments which makes us believe that this conjecture could be indeed true. A more general form of the conjecture would be: Any detector $A_m$ converging to a $m$-update local minima, would have an output same as the ML vector with high probability. It is easy to see from the analysis in Appendix C, that conjecture 1 stated for $A_1$ would only be true if the general conjecture were true for all $m = 2, 3, \cdots (2N_t - 1)$.

This therefore makes us believe that the general conjecture could also be true.

As shown in Appendix C, the validity of the conjecture rests on the validity of the fact that for any $1 \leq r < (2N_t - 1)$, if the detector output is a $r$-update local minima, then it is indeed a $r + 1$-update local minima with high probability. This is equivalent to the fact that the probabilities $p_{r,r+1}$ are high (close to 1) for all $1 \leq r < 2N_t - 1$. Due to the analytical difficulty involved in getting closed-form expressions for these probabilities, we evaluate them through Monte-Carlo simulations.
Figure 3.2: Conditional probabilities $p_{r,r+1}$ as a function of increasing $N_t = N_r$ for the 1-LAS detector with MMSE initial vector, 4-QAM, and SNR = 10 dB.

In Fig. 3.2, we plot the probabilities $p_{r,r+1}$ for the 1-LAS detector with $r = 1, 2, 3, 4$ and MMSE initial vector ('MMSE 1-LAS' detector), for increasing $N_t = N_r$, 4-QAM, and SNR = 10 dB. It is indeed observed that for a given $r$, $p_{r,r+1}$ initially decreases with increasing $N_r = N_t$, but eventually starts increasing for increasing $N_t = N_r$, and shows the tendency to converge to 1 in the limit as $N_t = N_r \to \infty$. In Fig. 3.3, these probabilities are shown for 'Random 1-LAS' detector (which is nothing but 1-LAS detector with any random data vector as the initial vector). Observations similar to those made in Fig. 3.2 for MMSE 1-LAS detector can be made for Random 1-LAS in Fig. 3.3; i.e., the probabilities for 1-LAS with any starting vector tend to 1 for large $N_t = N_r$. Also, it can be seen that for a given $N_t = N_r$ the probability values are smaller with Random 1-LAS compared to MMSE 1-LAS, indicating a faster convergence of MMSE 1-LAS compared to Random 1-LAS. The observations in Figs. 3.2 and 3.3 strengthen the validity of the conjecture 1, which is stated for any arbitrary detector converging to a 1-update local minima, and any initial data vector. □

Finally, we present the following conjecture on the error probability of 1-LAS detector and analyze it.
Figure 3.3: Conditional probabilities $p_{r,r+1}$ as a function of increasing $N_t = N_r$ for the 1-LAS detector with Random initial vector, 4-QAM and SNR = 10 dB.

**Conjecture 2.** The data vector/bit error probability of the 1-LAS detector converges to that of the ML detector as $N_t, N_r \to \infty$ with $N_t = N_r$.

Let $d_{LAS}$ be the final output symbol vector of the 1-LAS algorithm given $x$, $H$ and $n$. The algorithm satisfies property (3.13), and therefore conjecture 1 is applicable. With $A_1(.)$ as the 1-LAS algorithm, $p(x, n)$ can now be expressed as

$$p(x, n) = \mathbb{E}_H[I(n \in \mathcal{R}_d \mid H, x, n, d = A_1(H, y))]$$

$$= \mathbb{E}_H[I(n \in \mathcal{R}_d \mid H, x, n, d = \text{LAS}(H, y) = d_{LAS})]$$

$$= \mathbb{E}_H[I(n \in \mathcal{R}_{d_{LAS}} \mid H, x, n, d = \text{LAS}(H, y) = d_{LAS})]$$

$$= \mathbb{E}_H[I(d_{LAS} = \text{ML vector} \mid H, x, n, d = \text{LAS}(H, y) = d_{LAS})]$$

$$= \mathbb{E}_H[I(d_{LAS} = \text{ML vector} \mid H, x, n)] \quad (3.17)$$

In the above derivation, we have used Lemma 3, which states that if $n \in \mathcal{R}_{d_{LAS}}$, then $d_{LAS}$ is indeed the ML vector for the given $x$, $H$ and $n$. 


Let us further define

\[ P(d_{LAS} = \text{ML vector} | x, n) \triangleq \mathbb{E}_H[I(d_{LAS} = \text{ML vector} | H, x, n)] \]

\[ P(d_{LAS} \neq \text{ML vector} | x, n) \triangleq 1 - E_H[I(d_{LAS} = \text{ML vector} | H, x, n)]. \] (3.18)

Assuming conjecture 1 to be true, we can state that for any \( \delta, 0 \leq \delta \leq 1 \), there exists an integer \( N(\delta) \) such that for any \( N_t \geq N(\delta) \), and any \((x, n)\),

\[ P(d_{LAS} = \text{ML vector} | x, n) > (1 - \delta). \] (3.19)

For any \( N_t \geq N(\delta) \) and a given \((x, n)\), the conditional probability of symbol vector error averaged over the distribution of \( H \) is given by

\[ P_{LAS}(\text{error} | x, n) \triangleq \mathbb{E}_H[I(d_{LAS} \neq x | H, x, n)] \]

\[ = \mathbb{E}_H[I(d_{LAS} \neq x, d_{LAS} = \text{ML vector} | H, x, n)] \]

\[ + \mathbb{E}_H[I(d_{LAS} \neq x, d_{LAS} \neq \text{ML vector} | H, x, n)] \]

\[ = P(d_{LAS} \neq x, d_{LAS} = \text{ML vector} | x, n) \]

\[ + P(d_{LAS} \neq x, d_{LAS} \neq \text{ML vector} | x, n) \]

\[ = P(d_{LAS} \neq x | d_{LAS} = \text{ML vector}, x, n)P(d_{LAS} = \text{ML vector} | x, n) \]

\[ + P(d_{LAS} \neq x | d_{LAS} \neq \text{ML vector}, x, n)P(d_{LAS} \neq \text{ML vector} | x, n). \] (3.20)

From (3.19), we have \( P(d_{LAS} \neq \text{ML vector} | x, n) \leq \delta \). Also, \( P(d_{LAS} \neq x | d_{LAS} = \text{ML vector}, x, n) \) can be bounded as follows. Firstly, using (3.19) we have

\[ P(d_{LAS} \neq x | d_{LAS} = \text{ML vector}, x, n) = \frac{P(d_{LAS} \neq x, d_{LAS} = \text{ML vector} | x, n)}{P(d_{LAS} = \text{ML vector} | x, n)} \]

\[ < \frac{P(d_{LAS} \neq x, d_{LAS} = \text{ML vector} | x, n)}{(1 - \delta)}. \] (3.21)
\[ P(d_{LAS} \neq x, d_{LAS} = \text{ML vector} \mid x, n) \] can be expressed as
\[
P(d_{LAS} \neq x, d_{LAS} = \text{ML vector} \mid x, n) = P(x \neq \text{ML vector}, d_{LAS} = \text{ML vector} \mid x, n) \\
= P(x \neq \text{ML vector} \mid x, n) \\
- P(x \neq \text{ML vector}, d_{LAS} \neq \text{ML vector} \mid x, n) \\
\leq P(x \neq \text{ML vector} \mid x, n). \tag{3.22}
\]

\(P(x \neq \text{ML vector} \mid x, n)\) is nothing but the probability of error of the ML detector (averaged over the distribution of H) for a given \(x, n\). Subsequently, we denote this by \(P_{ML(\text{error} \mid x, n)}\). Using (3.21) and (3.22), we have
\[
P(d_{LAS} \neq x \mid d_{LAS} = \text{ML vector}, x, n) < \frac{P_{ML(\text{error} \mid x, n)}}{(1 - \delta)}. \tag{3.23}
\]

Using (3.23) in (3.20), \(P_{LAS(\text{error} \mid x, n)}\) can be upper bounded as follows
\[
P_{LAS(\text{error} \mid x, n)} < \frac{P_{ML(\text{error} \mid x, n)}}{(1 - \delta)} + \delta P(d_{LAS} \neq x \mid d_{LAS} \neq \text{ML vector}, x, n) \\
\leq \frac{P_{ML(\text{error} \mid x, n)}}{(1 - \delta)} + \delta. \tag{3.24}
\]

Since the ML detector is the optimal detector w.r.t the symbol vector error probability, we have
\[
\frac{P_{ML(\text{error} \mid x, n)}}{(1 - \delta)} + \delta > P_{LAS(\text{error} \mid x, n)} > P_{ML(\text{error} \mid x, n)}. \tag{3.25}
\]

Let
\[
M(x) \triangleq \frac{1}{2} \left( (x + 2) - \sqrt{x^2 + 4} \right), \quad x > 0. \tag{3.26}
\]

\(M(x)\) is a monotonically increasing function of \(x\), which takes on positive values and is bounded above by 1. Therefore, for a given \(1 > \delta > 0\), there exists some \(x = \epsilon\) such
that \( M(\epsilon) = \delta \). Also \( M \) is an invertible function, and therefore it follows that

\[
\epsilon = \delta \left( 1 + \frac{1}{(1 - \delta)} \right). \tag{3.27}
\]

Using (3.25) and (3.27), the absolute difference between the error probability of the LAS and the ML detector can be written as

\[
|P_{LAS}(\text{error} | x, n) - P_{ML}(\text{error} | x, n)| = \left( P_{LAS}(\text{error} | x, n) - P_{ML}(\text{error} | x, n) \right)
\leq \frac{P_{ML}(\text{error} | x, n)}{(1 - \delta)} + \delta - P_{ML}(\text{error} | x, n)
= \delta \left( 1 + \frac{P_{ML}(\text{error} | x, n)}{(1 - \delta)} \right)
\leq \delta \left( 1 + \frac{1}{(1 - \delta)} \right) = \epsilon. \tag{3.28}
\]

We next define the averaged LAS and ML error probabilities as

\[
P_{LAS}(\text{error}) \triangleq \mathbb{E}_{x, n}[P_{LAS}(\text{error} | x, n)] \tag{3.29}
\]

\[
P_{ML}(\text{error}) \triangleq \mathbb{E}_{x, n}[P_{ML}(\text{error} | x, n)].
\]

From (3.28), the bound on the absolute difference \( |P_{LAS}(\text{error} | x, n) - P_{ML}(\text{error} | x, n)| \) is independent of \((x, n)\), and therefore averaging over \((x, n)\) results in

\[
|P_{LAS}(\text{error}) - P_{ML}(\text{error})| < \epsilon. \tag{3.30}
\]

Hence, we have shown that for any arbitrary \( \epsilon > 0 \), there exists a corresponding positive real \( \delta = M(\epsilon) \) and an integer \( F(\epsilon) \triangleq N(M(\epsilon)) = N(\delta) \) such that, for all \( N_t \geq F(\epsilon), |P_{LAS}(\text{error}) - P_{ML}(\text{error})| < \epsilon \). This proves that \( P_{LAS}(\text{error}) \to P_{ML}(\text{error}) \) as \( N_t \to \infty \). This analysis can be adapted to show that, apart from the symbol vector error probability, the bit error probability of the LAS detector also converges to that of the ML detector. The analysis for the bit error probability convergence is along similar lines, except that instead of defining the error event as \( d_{LAS} \neq x \), we define error events
for each bit. For example, for the $p$th bit, the error event is defined as $d_{p \text{LAS}} \neq x_p$. □

### 3.2 Simulation Results and Discussions

In Fig. 3.4, we show the simulated BER performance of the 1-LAS detector with MMSE initial vector for V-BLAST MIMO as a function of SNR for increasing $N_t = N_r$. The modulation alphabet is 4-QAM. Since an analytical expression for ML performance in the large MIMO system limit is not available, and simulating the ML performance for large dimensions involves prohibitively high complexity, we plot the SISO AWGN performance as a lower bound for comparison. It can be seen that for increasing $N_t = N_r$, the BER performance of the 1-LAS detector approaches the SISO AWGN performance at high SNRs, which lends simulation support for conjecture 2.

In Figs. 3.5 and 3.6, we plot the simulated BER performance of the 1-LAS detector for
Figure 3.5: Simulated BER performance of 1-LAS detector with Random initial vector for increasing $N_t = N_r$. 4-QAM.

Figure 3.6: Simulated BER performance of the 1-LAS detector with MF initial vector for increasing $N_t = N_r$. 4-QAM.
random initial vector and matched filter (MF) initial vector\(^1\), respectively, for increasing \(N_r = N_t\) and 4-QAM. It is seen that, regardless on the initial vector used, the 1-LAS algorithm performance tends towards ML performance for large \(N_t = N_r\). Therefore, the asymptotic convergence of the LAS algorithm appears to be invariant to the choice of initial data vector. Note that conjecture 2 does not assume anything about the initial data vector \(d^{(0)}\). Also, comparing the performance with MMSE and Random initial vectors in Figs. 3.4 and 3.5, we see that a good initial vector (e.g., MMSE initial vector) allows a faster convergence than a poor initial vector (e.g., random initial vector).

We note that there exists a relation between the error performance and the probabilities \(p_{r,r+1}\). Instead of making observations from the error performance curves, we can also infer many properties and behavior of the LAS detector from the probabilities \(p_{r,r+1}\).

As an example, from Figs. 3.2 and 3.3, we note that the product \(p_{1,2}p_{2,3}\) is much smaller for the Random 1-LAS detector than for the MMSE 1-LAS detector. This suggests that the relative performance improvement in going from Random 1-LAS to Random 3-LAS should be more than the performance improvement in going from MMSE 1-LAS

\(^1\)MF initial vector is given by \(\text{sgn}(H^T y)\).
to MMSE 3-LAS. From the simulated error performance comparison between MMSE 1-LAS and MMSE 3-LAS presented in Chapter 2, and a similar comparison in Fig. 3.7, we observe that, indeed, the relative performance improvement from 1-LAS to 3-LAS is more for Random initial vector.
Chapter 4

Large-MIMO Detection Using Probabilistic Data Association

In the previous two chapters, we dealt with LAS algorithm, which is a local neighborhood search algorithm, and demonstrated its suitability for large-MIMO detection. In this chapter, we present another low-complexity algorithm suited for large-MIMO detection. The algorithm is based on probabilistic data association (PDA), which was originally developed for target tracking, and is widely being employed in digital communications [60]-[66]. PDA algorithm is a reduced complexity alternative to the a posteriori probability (APP) decoder/detector/equalizer. In this chapter, we develop the PDA algorithm for detection of MIMO signals, and demonstrate its suitability for large-MIMO detection in terms of both performance and complexity [83].

4.1 System Model

We assume a MIMO channel with $N_t$ transmit and $N_r$ receive antennas. The $N_tN_r$ channel gains are modeled as i.i.d. $\mathcal{CN}(0,1)$. The channel gains are assumed to be known at the receiver but not at the transmitter. We consider two types of MIMO transmit architectures, namely, V-BLAST and STBC MIMO. As seen in Chapter 2, the
following equivalent real valued system model is applicable for both architectures:

\[
y = H'x + n,
\]

where \( H' \in \mathbb{R}^{2N_r p \times 2k} \) is the equivalent channel matrix, \( y \in \mathbb{R}^{2N_r p} \) is the equivalent received vector, \( x \in \mathbb{A}^{2k} \) is the vector of transmitted symbols, and \( n \in \mathbb{R}^{2N_r p} \) is the noise vector. For a V-BLAST MIMO system with \( N_r \) receive and \( N_t \) transmit antennas, we have \( p = 1 \) and \( k = N_t \) in the above model. For a MIMO system using STBCs from CDA [21], we have \( p = N_t \) and \( k = N_t^2 \).

The entries of the noise vector \( n \) are modeled as i.i.d \( \mathcal{N}(0, \sigma^2 = \frac{N_t E_s^2}{2\gamma}) \), where \( E_s \) is the average energy of the transmitted complex symbols, and \( \gamma \) is the average received SNR per receive antenna. Each element of \( x \) is a \( \sqrt{M} \)-PAM symbol. \( \sqrt{M} \)-PAM symbols take discrete values from \( \mathbb{A} \triangleq \{ a_q, q = 1, \cdots, \sqrt{M} \} \), where \( a_q = \sqrt{\frac{3E_s}{2(M-1)}(2q-1-\sqrt{M})} \). The STBCs considered are non-orthogonal STBCs from CDA [21].

### 4.2 Proposed PDA Based Detection

In this section, we present the proposed PDA based detection algorithm. In the real-valued system model in (4.1), each entry of \( x \) belongs to a \( \sqrt{M} \)-PAM constellation, where \( M \) is the size of the original square QAM constellation of the transmitted complex symbols. Let \( b_i^{(0)}, b_i^{(1)}, \cdots, b_i^{(q-1)} \) denote the \( q = \log_2(\sqrt{M}) \) constituent bits of the \( i \)th entry \( x_i \) of \( x \). We can write the value of each entry of \( x \) as a linear combination of its constituent bits as

\[
x_i = \sum_{j=0}^{q-1} 2^j b_i^{(j)}, \quad i = 0, 1, \cdots, 2k - 1.
\]

\(^{1}\)We consider square QAM modulation. Nevertheless, applicability of the algorithm to rectangular QAM is straightforward.
Let \( b \in \{+1, -1\}^{2qk} \), defined as

\[
b \triangleq \begin{bmatrix} b_0^{(0)} & \cdots & b_0^{(q-1)} & b_1^{(0)} & \cdots & b_1^{(q-1)} & \vdots & \vdots & \cdots & \vdots & b_{2k-1}^{(0)} & \cdots & b_{2k-1}^{(q-1)} \end{bmatrix}^T,
\]

(4.3)
denote the transmitted bit vector. Defining \( c \triangleq [2^0 2^1 \cdots 2^{q-1}] \), we can write \( x \) as

\[
x = (I_{2k} \otimes c)b,
\]

(4.4)
using which we can rewrite (4.1) as

\[
y = H'(I_{2k} \otimes c)b + n, \quad \triangleq H
\]

(4.5)
where \( H \in \mathbb{R}^{2N_r p \times 2qk} \) is the effective channel matrix. The MAP estimate of bit \( b_i^{(j)} \) is

\[
\hat{b}_i^{(j)} = \arg \max_{a \in \{\pm 1\}} p(b_i^{(j)} = a | y, H),
\]

(4.6)
whose computational complexity is exponential in \( k \). Our goal is to obtain \( \hat{b} \), an estimate of \( b \), at low complexities. For this, we iteratively update the statistics of each bit of \( b \), as described in the following subsection, for a certain number of iterations, and hard decisions are made on the final statistics to get \( \hat{b} \).

### 4.2.1 Iterative Procedure

The algorithm is iterative in nature, where \( 2qk \) statistic updates, one for each of the constituent bits, are performed in each iteration. We start the algorithm by initializing the a priori probabilities as \( P(b_i^{(j)} = +1) = P(b_i^{(j)} = -1) = 0.5, \forall i = 0, \cdots, 2k - 1 \) and \( j = 0, \cdots, q-1 \). In an iteration, the statistics of the bits are updated sequentially, i.e., the ordered sequence of updates in an iteration is \{ \( b_0^{(0)}, \cdots, b_0^{(q-1)}, \cdots, b_{2k-1}^{(0)}, \cdots, b_{2k-1}^{(q-1)} \} \). The steps involved in each iteration of the algorithm are derived as follows.
The likelihood ratio of bit $b_i^{(j)}$ in an iteration, denoted by $\Lambda_i^{(j)}$, is given by

$$
\Lambda_i^{(j)} = \frac{P(y | b_i^{(j)} = +1)}{P(y | b_i^{(j)} = -1)} \frac{P(b_i^{(j)} = +1)}{P(b_i^{(j)} = -1)}.
$$

Denoting the $t$th column of $H$ by $h_t$, we can write (4.5) as

$$
y = h_{qi+j} b_i^{(j)} + \sum_{l=0}^{2k-1} \sum_{m=0}^{q-1} h_{ql+m} b_l^{(m)} + n,
$$

where $n \in \mathbb{R}^{2N_t p}$ is the interference plus noise vector. To calculate $\beta_i^{(j)}$, we approximate the distribution of $n$ to be Gaussian, and hence $y$ is Gaussian conditioned on $b_i^{(j)}$. Since there are $2qk - 1$ terms in the double summation in (4.8), this Gaussian approximation gets increasingly accurate for large $k$ and hence for large $N_t$ (note that $k$ is proportional to $N_t$). Since a Gaussian distribution is fully characterized by its mean and covariance, we evaluate the mean and covariance of $y$ given $b_i^{(j)} = +1$ and $b_i^{(j)} = -1$. For notational simplicity, let us define $p_i^{j+} \triangleq P(b_i^{(j)} = +1)$ and $p_i^{j-} \triangleq P(b_i^{(j)} = -1)$, where $p_i^{j+} + p_i^{j-} = 1$. Let $\mu_i^{j+} \triangleq \mathbb{E}(y | b_i^{(j)} = +1)$ and $\mu_i^{j-} \triangleq \mathbb{E}(y | b_i^{(j)} = -1)$. Now, from (4.8), we can write $\mu_i^{j+}$ as

$$
\mu_i^{j+} = h_{qi+j} + \sum_{l=0}^{2k-1} \sum_{m=0}^{q-1} h_{ql+m} (2p_i^{m+} - 1).
$$

Similarly, we can write $\mu_i^{j-}$ as

$$
\mu_i^{j-} = -h_{qi+j} + \sum_{l=0}^{2k-1} \sum_{m=0}^{q-1} h_{ql+m} (2p_i^{m+} - 1)
$$

$$
= \mu_i^{j+} - 2h_{qi+j}.
$$
Next, the $2N_rp \times 2N_rp$ covariance matrix, $C_i^j$, of $y$ given $b_i^j$ is

$$C_i^j = \mathbb{E}\left\{ \left[ n + \sum_{l=0}^{2k-1} \sum_{m=0}^{q-1} h_{ql+m}(b_i^{(m)} - 2p_i^{m+} + 1) \right] \cdot \left[ n + \sum_{l=0}^{2k-1} \sum_{m=0}^{q-1} h_{ql+m}(b_i^{(m)} - 2p_i^{m+} + 1) \right]^T \right\}. \quad (4.11)$$

Assuming independence among the constituent bits, we can simplify $C_i^j$ in (4.11) as

$$C_i^j = \sigma^2 I_{2N_rp} + \sum_{l=0}^{2k-1} \sum_{m=0}^{q-1} \sum_{m \neq q(l-1)+j} h_{ql+m} h_{ql+m}^T 4p_i^{m+}(1-p_i^{m+}). \quad (4.12)$$

Using the above mean and covariance, we can write the distribution of $y$ given $b_i^{(j)} = \pm 1$ as

$$P(y|b_i^{(j)} = \pm 1) = e^{-\frac{(y-\mu_i^{(j)})^T(C_i^j)^{-1}(y-\mu_i^{(j)})}{(2\pi)^{N_rp}|C_i^j|^\frac{1}{2}}}. \quad (4.13)$$

Using (4.13), $\beta_i^j$ can be written as

$$\beta_i^j = e^{-\frac{(y-\mu_i^+)^T(C_i^j)^{-1}(y-\mu_i^+)-(y-\mu_i^-)^T(C_i^j)^{-1}(y-\mu_i^-)}{2}}. \quad (4.14)$$

Using $\alpha_i^{(j)}$ and $\beta_i^{(j)}$, $\Lambda_i^{(j)}$ is computed using (4.7). Using the value of $\Lambda_i^{(j)}$, and using $P(b_i^{(j)} = +1|y) + P(b_i^{(j)} = -1|y) = 1$, the statistics of $b_i^{(j)}$ is updated as

$$P(b_i^{(j)} = +1|y) = \frac{\Lambda_i^{(j)}}{1 + \Lambda_i^{(j)}}, \quad P(b_i^{(j)} = -1|y) = \frac{1}{1 + \Lambda_i^{(j)}}. \quad (4.15)$$

This completes one iteration of the algorithm; i.e., each iteration involves the computation of $\alpha_i^{(j)}$ and (4.9), (4.10), (4.12), (4.14), (4.7), and (4.15) for all $i, j$. The updated values of $P(b_i^{(j)} = +1|y)$ and $P(b_i^{(j)} = -1|y)$ in (4.15) for all $i, j$ are fed back as a priori probabilities to the next iteration. The algorithm terminates after a certain number of such iterations. At the end of the last iteration, decision is made on the final statistics
to obtain the bit estimate $\hat{b}_i^{(j)}$ as $+1$ if $\Lambda_i^{(j)} \geq 1$, and $-1$ otherwise. In coded systems, $\Lambda_i^{(j)}$'s are fed as soft inputs to the decoder.

### 4.2.2 Complexity Reduction

The most computationally expensive operation in computing $\beta_i^{(j)}$ is the evaluation of the inverse of the covariance matrix, $C_{i}^{j}$, of size $2N_r \times 2N_r$ which requires $O(N_r^3 p^3)$ complexity, which can be reduced as follows. Define matrix $D$ as

$$D \triangleq \sigma^2 I_{2N_r p} + \sum_{l=0}^{2k-1} \sum_{m=0}^{q-1} h_{ql+m} h_{ql+m}^T 4p_{i}^{m+} (1 - p_{i}^{m+}).$$  \hspace{1cm} (4.16)

At the start of the algorithm, with $p_{i}^{j+} = p_{i}^{j-} = 0.5$, $\forall i, j$, $D$ becomes $\sigma^2 I_{2N_r p} + HH^T$.

**Computation of $D^{-1}$:** When the statistics of $b_{i}^{(j)}$ is updated using (4.15), the $D$ matrix in (4.16) also changes. Inversion of this updated $D$ would require $O(N_r^3 p^3)$ complexity. However, $D^{-1}$ can be obtained from the previously available $D^{-1}$ in $O(N_r^2 p^2)$ complexity as follows. Since the statistics of only $b_{i}^{(j)}$ is updated, the new $D$ is just a rank one update of the old $D$. So, using the matrix inversion lemma, the new $D^{-1}$ can be obtained from the old $D^{-1}$ as

$$D^{-1} \leftarrow D^{-1} - D^{-1} h_{ni+j} h_{ni+j}^T D^{-1} \frac{1}{\eta},$$  \hspace{1cm} (4.17)

where $\eta = 4p_{i}^{j+} (1 - p_{i}^{j+}) - 4p_{i,old}^{j+} (1 - p_{i,old}^{j+})$, and $p_{i}^{j+}$ and $p_{i,old}^{j+}$ are the new (i.e., after the update in (4.15)) and old (before the update) values, respectively. Both the numerator and denominator in the 2nd term on the RHS of (4.17) can be computed in $O(N_r^2 p^2)$ complexity. So, the computation of the new $D^{-1}$ using the old $D^{-1}$ can be done in $O(N_r^2 p^2)$ complexity.

**Computation of $(C_{i}^{j})^{-1}$:** Using (4.16) and (4.12), we can write $C_{i}^{j}$ in terms of $D$ as

$$C_{i}^{j} = D - 4p_{i}^{j+} (1 - p_{i}^{j+}) h_{qi+j} h_{qi+j}^T.$$  \hspace{1cm} (4.18)
We can compute \((C_j^i)^{-1}\) from \(D^{-1}\) at reduced complexity using the matrix inversion lemma, as

\[
(C_j^i)^{-1} = D^{-1} - \frac{D^{-1} h_{qi+j} h_{qi+j}^T D^{-1}}{h_{qi+j}^T D^{-1} h_{qi+j} - \frac{1}{4p_j^t (1-p_j^t)}},
\]

which can be computed in \(O(N_t^2 p^2)\) complexity.

**Computation of \(\mu_j^i\):** Computation of \(\beta_j^i\) involves the computation of \(\mu_j^i\) and \(\mu_j^i\) also. From (4.10), it is clear that \(\mu_j^i\) can be computed from \(\mu_j^i\) with a computational overhead of only \(O(N_r p)\). From (4.9), it is seen that computing \(\mu_j^i\) would require \(O(q N_r pk)\) complexity. However, this complexity can be reduced as follows.

Define vector \(u\) as

\[
u \triangleq \sum_{i=0}^{2k-1} \sum_{m=0}^{q-1} h_{qi+m}(2p_i^{m+} - 1).
\]

Using (4.9) and (4.20), we can write \(\mu_j^i = u + 2(1 - p_j^i) h_{qi+j}\). \(u\) can be computed iteratively at \(O(N_r p)\) complexity as follows. When the statistics of \(h_j^{(i)}\) is updated, we can obtain the new \(u\) from the old \(u\) as \(u \leftarrow u + 2(p_i^{j+} - p_i^{j+}) h_{qi+j}\), whose complexity is \(O(N_r p)\). So, the computation of \(\mu_j^i\) and \(\mu_j^i\) needs \(O(N_r p)\) complexity. The full listing of the proposed algorithm is presented in Table-4.1.

**Overall Complexity:** We need to compute \(HH^T\) at the start of the algorithm. This requires \(O(qk N_r^2 p^2)\) complexity. So the computation of the initial \(D^{-1}\) requires \(O(qk N_r^2 p^2) + O(N_t^3 p^3)\). Based on the complexity reduction in Section 4.2.2, the complexity in updating the statistics of one constituent bit is \(O(N_r^2 p^2)\). So, the complexity for the update of all the \(2qk\) constituent bits in an iteration is \(O(qk N_r^2 p^2)\). Since the number of iterations is fixed, the overall complexity of the algorithm is \(O(qk N_r^2 p^2) + O(N_t^3 p^3)\). Note that, for a \(N_t = N_r\) V-BLAST MIMO system, since \(p = 1\), the per bit complexity is \(O(N_r^2)\), which is the same as that of the LAS algorithm. However, in case of STBC, since \(p = N_t\), PDA has a per bit complexity of \(O(N_t^4)\), which is an order higher than that of LAS.
Chapter 4. Large-MIMO Detection Using Probabilistic Data Association

Table 4.1: Listing of the proposed PDA based detection algorithm.

Initialisation
1. \( p_i^{+0} = p_i^{-0} = 0.5, \quad \Lambda_{i}^{(j)} = 1, \quad \forall i = 0, 1, \ldots, 2k - 1, j = 0, 1, \ldots, q - 1 \)
2. \( u = 0, \quad D^{-1} = (HH^T + \sigma^2 I)^{-1} \)
3. \( \text{num\_iter}: \) number of iterations
4. \( \kappa = 1; \quad \kappa \) is the iteration number

Statistics update in the \( \kappa \)th iteration
5. for \( i = 0 \) to \( 2k - 1 \)
6. for \( j = 0 \) to \( q - 1 \)

Update of statistics of bit \( b_i^{(j)} \)
7. \( \mu_i^{+j} = u + 2(1 - p_i^{+j}) h_{qi+j} \)
8. \( \mu_i^{-j} = \mu_i^{+j} - 2h_{qi+j} \)
9. \( (C_{i}^{j})^{-1} = D^{-1} - \frac{h_{qi+j}^T h_{qi+j} D^{-1}}{q_i^{+j}(1 - q_i^{+j})} \)
10. \( \beta_i^{j} = e^{-((y - \mu_i^{+j})^T (C_{i}^{j})^{-1}(y - \mu_i^{+j}) - (y - \mu_i^{-j})^T (C_{i}^{j})^{-1}(y - \mu_i^{-j}))} \)
11. \( p_{i,\text{odd}} = p_i^{+j}, \quad p_{i,\text{even}} = p_i^{-j} \)
12. \( \alpha_i^{(j)} = \frac{p_{i,\text{odd}}}{p_{i,\text{even}}} \)
13. \( \Lambda_i^{(j)} = \frac{\beta_i^{(j)} \alpha_i^{(j)}}{1 + \Lambda_i^{(j)}} \)
14. \( p_i^{+j} = \frac{\Lambda_i^{(j)}}{1 + \Lambda_i^{(j)}}, \quad p_i^{-j} = \frac{1}{1 + \Lambda_i^{(j)}} \)

Update of \( u \) and \( D^{-1} \)
15. \( u \leftarrow u + 2(p_i^{+j} - p_{i,\text{even}}) h_{qi+j} \)
16. \( \eta = 4p_i^{+j}(1 - p_i^{+j}) - 4p_{i,\text{even}}(1 - p_i^{+j}) \)
17. \( D^{-1} \leftarrow D^{-1} - \frac{D^{-1} h_{qi+j}^T h_{qi+j} D^{-1}}{\eta} \)
18. end; \quad \text{End of for loop starting at line 5}
19. if (\( \kappa = \text{num\_iter} \)) goto line 21
20. \( \kappa = \kappa + 1, \) goto line 5
21. \( \hat{b}_i^{(j)} = \text{sgn}\left(\log(\Lambda_i^{(j)})\right), \quad \forall i = 0, 1, \ldots, 2k - 1, j = 0, 1, \ldots, q - 1 \)
22. \( \hat{x}_i = \sum_{j=0}^{q-1} 2^j \hat{b}_i^{(j)}, \quad \forall i = 0, 1, \ldots, 2k - 1 \)
23. Terminate
4.3 Results and Discussions

In this section, we present the simulated BER performance of the proposed PDA based algorithm for detection in large V-BLAST and non-orthogonal STBC MIMO systems.

4.3.1 BER Performance in Large V-BLAST MIMO

BER performance with increasing number of PDA iterations: In Fig. 4.1, we plot the variation of the BER performance in a $N_r = N_t = 64$ V-BLAST MIMO system with 4-QAM, for increasing number of PDA iterations ($m = 1, 2, 4, 8$). Perfect CSIR is assumed at the receiver. It is observed that, as expected, the error performance improves with increase in the number of iterations. However, the performance improvement for more than 4 iterations is observed to be marginal.

Large-System Behavior of PDA Based Detection: In Fig. 4.2, we plot the BER performance of the PDA detector in V-BLAST MIMO with increasing number of transmit and receive antennas ($N_t = N_r = 8, 16, 32, 64, 96$) with 4-QAM and $m = 5$ iterations. It is seen that the error performance improves with increasing $N_t = N_r$. This shows that the
PDA algorithm, like the LAS algorithm in Chapter 2, exhibits large-system behavior, making it suited for large-MIMO detection.

### 4.3.2 BER Performance in Large STBC MIMO

In this subsection, we present the simulated BER performance results for non-orthogonal STBC MIMO systems with PDA detection. The MIMO channel is assumed to be quasi static (i.e., fade remains constant for the duration of one STBC block, and i.i.d from one STBC block to another). The STBCs used are ILL-only non-orthogonal STBCs from CDA [21].

**PDA versus LAS performance with 4-QAM:** In Fig. 4.3, we plot the uncoded BER of the PDA algorithm as a function of average received SNR per receive antenna, $\gamma$, in decoding $4 \times 4, \ 8 \times 8, \ 16 \times 16$ ILL-only STBCs from CDA with $N_t = N_r$ and 4-QAM. Perfect CSIR and i.i.d fading are assumed. For the same settings, the performance of the 1-LAS algorithm with MMSE initial vector is also plotted for comparison. From Fig. 4.3, it is seen that i) as in V-BLAST MIMO, the PDA algorithm exhibits ‘large-system behavior’ in STBC MIMO as well, i.e., BER improves with increased number of
dimensions (i.e., with increased $N_t = N_r$), and approaches SISO AWGN performance for increasing $N_t = N_r$. For e.g., performance close to within about 1 dB from SISO AWGN performance is achieved at $10^{-3}$ uncoded BER in decoding $16 \times 16$ STBC from CDA having 512 real dimensions, and this illustrates the ability of the PDA algorithm to achieve very good performance at low complexities in large dimensions, and $ii)$ with 4-QAM, PDA and LAS algorithms achieve almost the same performance.

**PDA versus LAS performance with 16-QAM:** Figure 4.4 presents an uncoded BER comparison between PDA and LAS algorithms in decoding ILL-only STBCs from CDA with $N_t = N_r$ and 16-QAM under perfect CSIR and i.i.d fading. It can be seen that the PDA algorithm performs better at low SNRs than the LAS algorithm. For e.g., with $8 \times 8$ and $16 \times 16$ STBCs, at low SNRs (e.g., $< 25$ dB for $16 \times 16$ STBC), PDA algorithm performs better by about 2 dB compared to LAS algorithm at $10^{-2}$ uncoded BER.

**Turbo coded BER performance of PDA:** Figure 4.5 shows the rate-3/4 turbo coded BER of the PDA algorithm under perfect CSIR and i.i.d fading for $12 \times 12$ ILL STBC with $N_t = N_r = 12$ and 4-QAM, which corresponds to a spectral efficiency of 18 bps/Hz.
Figure 4.4: Comparison between the uncoded BER performance of PDA and LAS algorithms in decoding $4 \times 4$, $8 \times 8$, $16 \times 16$ ILL-only STBCs. $N_t = N_r$, 16-QAM.

Figure 4.5: Turbo coded BER performance of the PDA algorithm in decoding $12 \times 12$ ILL-only STBC with $N_t = N_r = 12$, 4-QAM, rate-3/4 turbo code, 18 bps/Hz and $m = 10$ for i) perfect CSIR, and ii) estimated CSIR using 2 iterations between PDA decoding/channel estimation.
The theoretical minimum SNR required to achieve 18 bps/Hz spectral efficiency on a $N_t = N_r = 12$ MIMO channel with perfect CSIR and i.i.d fading is 4.3 dB (obtained through simulation of the ergodic capacity formula [2]). From Fig. 4.5, it is seen that the PDA algorithm is able to achieve vertical fall in coded BER within just about 5 dB from theoretical minimum SNR, which is a good nearness to capacity performance.

**Iterative Channel Estimation/Detection:** We next relax the perfect CSIR assumption by considering a training based iterative channel estimation/PDA decoding scheme. Transmission is carried out in frames, where one $N_t \times N_t$ pilot matrix (for training purposes) followed by $N_d$ data STBC matrices are sent in each frame. One frame length, $T$, (taken to be the channel coherence time) is $T = (N_d+1)N_t$ channel uses. The proposed scheme works as follows: 

1. obtain an MMSE estimate of the channel matrix during the pilot phase,
2. use the estimated channel matrix to decode the data STBC matrices using PDA algorithm, and
3. iterate between channel estimation and PDA decoding for a certain number of times.

For the $12 \times 12$ ILL-only STBC from CDA, in addition to perfect CSIR performance, Fig. 4.5 also shows the performance with CSIR estimated using the proposed iterative channel estimation/decoding scheme for $N_d = 1$ and $N_d = 8$. Two iterations between channel estimation and PDA decoding are used. With $N_d = 8$ (corresponding to large coherence times, i.e., slow fading) the BER and bps/Hz with estimated CSIR get closer to those with perfect CSIR.

**Effect of Spatial Correlation:** In Figs. 4.3 to 4.5, we assumed i.i.d fading. But spatial correlation at transmit/receive antennas and the structure of scattering and propagation environment can affect the rank structure of the MIMO channel resulting in degraded performance. We finally relaxed the i.i.d. fading assumption by considering the correlated MIMO channel model in [8], which takes into account carrier frequency ($f_c$), spacing between antenna elements ($d_t, d_r$) distance between transmit and receive antennas ($R$), and scattering environment. In Fig. 4.6, we plot the BER of the PDA...
algorithm in decoding $12 \times 12$ ILL-only STBC from CDA with perfect CSIR in i) i.i.d. fading, and ii) correlated MIMO fading model in [8]. We observe that spatial correlation results in degradation of both coded and uncoded BER performance of the PDA detector. However, having more receive antennas for the same receive aperture, can mitigate the effects of spatial correlation. It is seen that, in terms of uncoded BER, by having the aperture fixed to 72 cm and increasing $N_r$ from 12 to 18 results in increased diversity order when compared to $N_r = 12$. Also, for the case of $N_r = 12$, the coded BER performance does not show a vertical fall, whereas for $N_r = 18$, a vertical fall close to theoretical minimum SNR is achieved.

In summary, the proposed PDA algorithm is found to be a good low-complexity algorithm suited for large-MIMO detection.
Chapter 5

Large-MIMO Precoding Using X- and Y-Codes

In this chapter, we consider precoding in large-MIMO systems, where CSI is perfectly available both at the transmitter and the receiver. We propose precoding schemes based on X- and Y-Codes to achieve high multiplexing and diversity gains at low complexity [92]-[94]. The proposed precoding schemes are based upon the singular value decomposition (SVD) of the channel matrix which transforms the MIMO channel into parallel subchannels. X- and Y-Codes are used to improve the diversity gain by pairing the subchannels, prior to SVD precoding. In particular, the subchannels with good diversity are paired with those having low diversity gains. Hence, a pair of channels is jointly encoded using a $2 \times 2$ real matrix, which is fixed a priori and does not change with each channel realization. For X-Codes these matrices are 2-dimensional rotation matrices parameterized by a single angle, while for Y-Codes, these matrices are 2-dimensional upper left triangular matrices. The complexity of the maximum likelihood decoding for both X- and Y-Codes is low. Specifically, the decoding complexity of Y-Codes is the same as that of a scalar channel. Moreover, we propose X-, Y-Precoders with the same structure as X-, Y-Codes, but the encoding matrices adapt to each channel realization. The optimal encoding matrices for X-,...
Y-Codes/Precoders are derived analytically. It is observed that X-Codes/Precoders perform better for well-conditioned channels, while Y-Codes/Precoders perform better for ill-conditioned channels, when compared to other precoding schemes in the literature.

This chapter is organized as follows. Section 5.1 introduces the system model and SVD precoding. In Section 5.2, we present the pairing of subchannels as a general coding strategy to achieve higher diversity order in fading channels. In Section 5.3, we propose the X-Codes and the X-Precoders. We show that ML decoding can be achieved with \( N_r \) 2-D real sphere decoders (SD). We also analyze the error performance and present the design of optimal X-Codes and X-Precoders. In Section 5.4, we propose the Y-Codes and Y-Precoder. We show that they have very low decoding complexity. We analyze the error performance and derive expressions for the optimal Y-Codes and Y-Precoders. Section 5.5 shows the simulation results and comparisons with other precoders. Section 5.6 discusses the complexity of the X-, Y-Codes/Precoders in comparison with other precoders.

### 5.1 System Model and SVD Precoding

We consider a \( N_t \times N_r \) MIMO system (\( N_r \leq N_t \)) with \( N_r \) receive and \( N_t \) transmit antennas. CSI is assumed to be known perfectly at both transmitter and receiver. Let \( x = (x_1, \cdots, x_{N_t})^T \) be the vector of symbols transmitted by the \( N_t \) transmit antennas, and let \( H = \{h_{ij}\}, i = 1, \cdots, N_r, j = 1, \cdots, N_t \), be the \( N_r \times N_t \) channel coefficient matrix, where \( h_{ij} \) is the complex channel gain between the \( j \)th transmit antenna and the \( i \)th receive antenna, and \( h_{ij} \)'s are modeled as i.i.d. and \( \mathcal{CN}(0, 1) \). The \( N_r \times 1 \) received vector is given by

\[
y = Hx + n, \tag{5.1}
\]
where $n$ is a spatially uncorrelated Gaussian noise vector such that $\mathbb{E}[nn^H] = N_0 I_{N_r}$. Such a system has a maximum multiplexing gain of $N_r$.

Let the number of transmitted information symbols be $N_s$ ($N_s \leq N_r$). The information bits are first mapped to the information symbol vector $u = (u_1, \cdots, u_{N_s})^T \in \mathbb{C}^{N_s}$, which is then mapped to the data symbol vector $z = (z_1, \cdots, z_{N_s})^T \in \mathbb{C}^{N_s}$ using a $N_s \times N_s$ matrix $G$ as

$$z = Gu + u^0,$$  \hspace{1cm} (5.2)

where $u^0 \in \mathbb{C}^{N_s}$ is a displacement vector used to reduce the average transmitted power. Let $T$ be the $N_t \times N_s$ precoding matrix which is applied to the data symbol vector to yield the transmitted vector

$$x = Tz.$$

In general, $T$, $G$, and $u_0$ are derived from the knowledge of $H$ at the transmitter and they are crucial to the system performance and complexity. The transmission power constraint is given by

$$\mathbb{E}[\|x\|^2] = P_T,$$  \hspace{1cm} (5.4)

where $P_T$ total transmit power, and we define the SNR as $\gamma \triangleq \frac{P_T}{N_0}$.

The proposed X- and Y-Codes are based on the SVD precoding technique, which is based on the singular value decomposition of the channel matrix $H = U \Lambda V$, where $U \in \mathbb{C}^{N_r \times N_r}$, $\Lambda \in \mathbb{C}^{N_r \times N_r}$, $V \in \mathbb{C}^{N_r \times N_t}$, $UU^H =VV^H = I_{N_r}$, and $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_{N_r})$, with $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_{N_r} \geq 0$. Let $\tilde{V} \in \mathbb{C}^{N_s \times N_t}$ be the submatrix with the first $N_s$ rows of
V. The SVD precoder uses

\[ T = \tilde{V}^H \]
\[ G = I_{N_s} \]
\[ u^0 = 0, \quad (5.5) \]

and the receiver gets

\[ y = HTu + n. \quad (5.6) \]

Let \( \tilde{U} \in \mathbb{C}^{N_r \times N_s} \) be the submatrix with the first \( N_s \) columns of \( U \). The receiver then computes

\[ r = \tilde{U}^H y = \tilde{\Lambda} u + w, \quad (5.7) \]

where \( w \in \mathbb{C}^{N_s} \) is still an uncorrelated Gaussian noise vector with \( \mathbb{E}[ww^H] = N_0 I_{N_s} \), \( \tilde{\Lambda} \triangleq \text{diag}(\lambda_1, \cdots, \lambda_{N_s}) \), and \( r = (r_1, \cdots, r_{N_s})^T \). SVD precoding therefore transforms the channel into \( N_s \) parallel channels

\[ r_i = \lambda_i u_i + w_i, \quad i = 1, \cdots, N_s, \quad (5.8) \]

with non-negative fading coefficients \( \lambda_i \). The overall error performance is dominated by the minimum singular value \( \lambda_{N_s} \). When \( N_s = N_r = N_t \), the resulting diversity order is only 1.

5.2 Pairing Good and Bad Subchannels

Without loss of generality, we consider only the full-rate SVD precoding scheme with even \( N_r \) and \( N_s = N_r \). The matrix \( G \in \mathbb{C}^{N_r \times N_r} \) is now used to pair (jointly encode)
different subchannels in order to improve the diversity order of the system. The precoding matrix $\mathbf{T} \in \mathbb{C}^{N_t \times N_r}$ and the transmitted vector $\mathbf{x}$ are given by

$$
\mathbf{T} = \mathbf{V}^H, \quad \mathbf{x} = \mathbf{V}^H(\mathbf{G}\mathbf{u} + \mathbf{u}^0).
$$

(5.9)

Let the list of subchannel pairings be $\{(i_k, j_k) \in [1, N_r] \times [1, N_r], \ k \in [1, N_r/2] \ | \ i_k < j_k\}$. An example of a list of pairings with $N_r = 6$ is $\{(1, 6), (2, 5), (3, 4)\}$. A subchannel can only be paired with some other channel exactly once. So $\{(1, 3), (1, 4), (2, 6)\}$ is not a valid pairing. On the $k$-th pair, consisting of subchannels $i_k$ and $j_k$, the information symbols $u_{i_k}$ and $u_{j_k}$ are jointly coded using a $2 \times 2$ matrix $\mathbf{A}_k$. In order to reduce the ML decoding complexity, we restrict the entries of $\mathbf{A}_k$ to be real valued. Each $\mathbf{A}_k \triangleq \{a_{k,i,j}\}$, $i, j \in [1, 2]$, is a submatrix of the code matrix $\mathbf{G}$ as shown below:

$$
g_{i_k,i_k} = a_{k,1,1} \quad g_{i_k,j_k} = a_{k,1,2} \\
g_{j_k,i_k} = a_{k,2,1} \quad g_{j_k,j_k} = a_{k,2,2}
$$

(5.10)

where $g_{i,j}$ is the entry of $\mathbf{G}$ in the $i$th row and $j$th column.

We shall see later, that an optimal pairing in terms of achieving the best diversity order is one in which the $k$th subchannel is paired with the $(N_r - k + 1)$th subchannel. For e.g., with $N_r = 6$, the X-Code structure is given by

$$
\mathbf{G} = \begin{bmatrix}
a_{1,1,1} & a_{1,1,2} \\
a_{2,1,1} & a_{2,1,2} \\
a_{3,1,1} & a_{3,1,2} \\
a_{3,2,1} & a_{3,2,2} \\
a_{2,2,1} & a_{2,2,2} \\
a_{1,2,1} & a_{1,2,2}
\end{bmatrix},
$$

(5.11)
and the Y-Code structure is given by\(^1\)

\[
\mathbf{G} = \begin{bmatrix}
a_{1,1,1} & a_{1,1,2} \\
a_{2,1,1} & a_{2,1,2} \\
a_{3,1,1} & a_{3,1,2} \\
a_{3,2,1} & a_{2,2,1} \\
a_{1,2,1}
\end{bmatrix}.
\tag{5.12}
\]

Let

\[
\mathbf{u}_k \triangleq [u_{i_k}, u_{j_k}]^T.
\tag{5.13}
\]

Due to the transmit power constraint in (5.4), and uniform power allocation between the \(N_r/2\) pairs, the encoder matrices \(\mathbf{A}_k\) must satisfy

\[
\mathbb{E}[\|\mathbf{A}_k \mathbf{u}_k + \mathbf{u}_k^0\|^2] = \frac{2P_T}{N_r}.
\tag{5.14}
\]

The expectation in (5.14) is over the distribution of the information symbol vector \(\mathbf{u}_k\).

\(\mathbf{u}_k^0\) is the subvector of the displacement vector \(\mathbf{u}^0\) for the \(k\)th pair.

The matrices \(\mathbf{A}_k\) for X- and Y- codes can be either fixed \textit{a priori} or can change with every channel realization. The latter case leads to the X- and Y-Precoders.

### 5.2.1 ML Decoding

Given the received vector \(\mathbf{y}\), the receiver computes

\[
\mathbf{r} = \mathbf{U}^H \mathbf{y} - \mathbf{A} \mathbf{u}^0.
\tag{5.15}
\]

\(^1\)The names X- and Y-Codes are due to the structure of the code generating matrices in 5.11 and 5.12.
Using (5.1) and (5.9), we can rewrite (5.15) as

\[ r = \Lambda G u + w = M u + w, \] (5.16)

where \( M \triangleq \Lambda G \) is the equivalent channel gain matrix and \( w \triangleq U^H n \) is a noise vector with the same statistics as \( n \). Further, we let

\[ r_k \triangleq [r_{i_k}, r_{j_k}]^T \]
\[ w_k \triangleq [w_{i_k}, w_{j_k}]^T. \]

Let \( M_k \in \mathbb{R}^{2 \times 2} \) denote the \( 2 \times 2 \) submatrix of \( M \) consisting of entries in the \( i_k \) and \( j_k \) rows and columns. Then (5.16) can be equivalently written as

\[ r_k = M_k u_k + w_k, \quad k = 1, \ldots, \frac{N_r}{2}. \] (5.17)

\( \mathcal{R}(u_k) \in \mathcal{S}_k \), where \( \mathcal{S}_k \) is a finite signal set in the 2-dimensional real space. Assuming that the same set is used for the imaginary component, the spectral efficiency, \( \eta \), is given by

\[ \eta = 2 \sum_{k=1}^{\frac{N_r}{2}} \log_2(|\mathcal{S}_k|). \] (5.18)

From (5.17), it is clear that the ML decoding (MLD) reduces to separate MLDs of the \( k \) pairs, which can be further separated into independent ML decoding of the real and imaginary components of \( u_k \). Then the MLD for the \( k \)-th pair is given by

\[ \mathcal{R}(\hat{u}_k) = \arg \min_{\mathcal{R}(u_k) \in \mathcal{S}_k} \| \mathcal{R}(r_k) - M_k \mathcal{R}(u_k) \|^2, \] (5.19)

and

\[ \mathcal{I}(\hat{u}_k) = \arg \min_{\mathcal{I}(u_k) \in \mathcal{S}_k} \| \mathcal{I}(r_k) - M_k \mathcal{I}(u_k) \|^2, \] (5.20)
where \( \hat{u}_k \) is the output of the ML detector for the \( k \)th pair.

### 5.2.2 Performance Analysis

Let \( P_k \) denote the word error probability (WEP) for the \( k \)th pair of subchannels with the ML receiver. The overall WEP for the transmitted information symbol vector is given by

\[
P = 1 - \prod_{k=1}^{N_r/2} (1 - P_k).
\] (5.21)

From (5.19) and (5.20), we see that WEPs for the real and the imaginary components of the \( k \)th pair are the same. Therefore, without loss of generality, we can compute the WEP only for the real component, denoted by \( P'_k \), and then \( P_k = 1 - (1 - P'_k)^2 \).

Let us further denote by \( P'_k(\Re(u_k)) \) the probability of the real part of the ML decoder decoding not in favor of \( \Re(u_k) \) when \( u_k \) is transmitted on the \( k \)th pair. \( P'_k \) can then be expressed in terms of \( P'_k(\Re(u_k)) \) as

\[
P'_k = \frac{1}{|S_k|} \sum_{\Re(u_k)} P'_k(\Re(u_k)),
\] (5.22)

where \( P'_k(\Re(u_k)) \) has to be evaluated differently for X-, Y-Codes and X-, Y-Preceders.

To explain this difference, we need the following definitions.

For a given channel realization, i.e., deterministic value of \( \lambda_{ik} \) and \( \lambda_{jk} \) for the \( k \)th pair, we let \( P'_k(\Re(u_k), \lambda_{ik}, \lambda_{jk}, A_k) \) to be the error probability of MLD for the real component of the \( k \)th channel, given that the information symbol \( u_k \) was transmitted on the \( k \)th pair. For X-, Y-Codes, the matrices \( A_k \) are fixed \textit{a priori} and are not function of the deterministic value of channel gains, and therefore \( P'_k(\Re(u_k)) \) is given by

\[
P'_k(\Re(u_k)) = \mathbb{E}_{(\lambda_{ik}, \lambda_{jk})} \left[ P'_k(\Re(u_k), \lambda_{ik}, \lambda_{jk}, A_k) \right].
\] (5.23)
We observe that $P'_k(\Re(u_k))$ is actually a function of $A_k$, and therefore the optimal error performance is obtained by minimizing (5.22) over $A_k$. Then the optimal matrix for the $k$th pair is given by

$$A_{k}^{opt} = \arg \min_{A_k} \sum_{\Re(u_k)} \mathbb{E}^{(\lambda_{ik}, \lambda_{jk})} \left[ P'_k(\Re(u_k), \lambda_{ik}, \lambda_{jk}, A_k) \right].$$  \hfill (5.24)

The minimization in (5.24) is constrained over matrices $A_k$ which satisfy (5.14). The optimal error performance $P_k^{opt}$ is given by

$$P_k^{opt} = \frac{1}{|S_k|} \sum_{\Re(u_k)} \mathbb{E}^{(\lambda_{ik}, \lambda_{jk})} \left[ P'_k(\Re(u_k), \lambda_{ik}, \lambda_{jk}, A_{k}^{opt}) \right].$$  \hfill (5.25)

For the X-, Y-Precoders, the matrices $A_k$ are chosen every time the channel changes. For optimal performance, the matrices $A_k$ are chosen so as to minimize the error probability for a given channel realization. $A_{k}^{opt}$, the optimal encoding matrix for the $k$th pair, is then given by

$$A_{k}^{opt}(\lambda_{ik}, \lambda_{jk}) = \arg \min_{A_k} \sum_{\Re(u_k)} P'_k(\Re(u_k), \lambda_{ik}, \lambda_{jk}, A_k).$$  \hfill (5.26)

The optimal error performance for X-, Y-Precoders is therefore given by

$$P_k^{opt} = \frac{1}{|S_k|} \sum_{\Re(u_k)} \mathbb{E}^{(\lambda_{ik}, \lambda_{jk})} \left[ P'_k(\Re(u_k), \lambda_{ik}, \lambda_{jk}, A_{k}^{opt}(\lambda_{ik}, \lambda_{jk})) \right].$$  \hfill (5.27)

Comparing (5.27) and (5.25), we immediately observe that the optimal error performance of X-, Y-Precoders is better than that of X-, Y-Codes.

Our next goal is to derive an analytic expression for $P'_k(\Re(u_k))$. We shall only discuss the derivation for X-, Y-Codes, since the performance of X-, Y-Precoders is better than that of X-, Y-Codes, and therefore have at least as much diversity order as X-, Y-Codes.

Getting an exact analytic expression for $P'_k$ is difficult, and so we try to get tight upper bounds using the union bound.
Theorem 1. The upper bound to $P'_k$ is given by

\[
P'_k \leq c_k(|S_k| - 1) \left( \frac{\gamma g_k(A_k)}{2P_T} \right)^{-\delta_k} + o(\gamma^{-\delta_k}),
\]

where

\[
\delta_k \triangleq (N_t - i_k + 1)(N_r - i_k + 1),
\]

\[
c_k \triangleq \frac{C(i_k)((2\delta_k - 1) \cdots 5 \cdot 3 \cdot 1)}{2\delta_k},
\]

and $g_k(A_k)$ is the generalized minimum distance, as defined in (D.5) (See Appendix D), $C(m)$ ($1 \leq m \leq \min(N_r, N_t)$) is defined in [101].

Proof: Proof of this theorem is given in Appendix D. 

Let us define the overall diversity order

\[
\delta_{ord} \triangleq \lim_{\gamma \to \infty} \frac{-\log P}{\log \gamma}.
\]

It is obvious that

\[
\delta_{ord} \geq \min_k \delta_k.
\]

This bound also holds for the X-, Y-Precoders since the error performance of the X-, Y-Precoders is always better than that of the X-, Y-Codes.

5.2.3 Design of Optimal Pairing

From the lower bound on $\delta_{ord}$ given by (5.31), it is clear that the following pairing of sub-channels

\[
i_k = k, \ j_k = (N_r - k + 1)
\]
achieves the following best lower bound

\[ \delta_{\text{ord}} \geq \left( \frac{N_r}{2} + 1 \right) \left( N_t - \frac{N_r}{2} + 1 \right). \]  

(5.33)

**Remark 1.** Note that this corresponds to a cross-form generator matrix \( G \), and is not the only pairing for the best lower bound. Also, we note that the diversity order improves significantly, when compared to the case of no pairing. It can be shown that if only \( N_s \) (\( N_s \) is even) out of the \( N_r \) subchannels are used for transmission, the lower bound on the achievable diversity order is \( (N_r - \frac{N_s}{2} + 1)(N_t - \frac{N_s}{2} + 1) \).

Although it is hard to compute \( A_{k,\text{opt}} \), we can compute the best \( A_{k,\star} \), denoted by \( A_{k,\star} \), which minimizes the upper bound on \( P_k' \) in (5.28). Then we have

\[
A_{k,\star} = \arg \max_{A_k} \mathbb{E} \left[ \| A_k u_k + u_0^k \|^2 \right] = \frac{g_k(A_k)}{2P_T}.
\]

(5.34)

Using (5.28), (5.32) and (5.34), we obtain

\[
P_k' \leq c_k(|S_k| - 1) \left( \frac{g_k(A_{k,\star})}{2P_T} \right)^{-\delta_k} + o(\gamma^{-\delta_k}),
\]

(5.35)

where \( \delta_k \overset{\Delta}{=} (N_t - k + 1)(N_r - k + 1) \).

### 5.3 X-Codes and X-Precoder

#### 5.3.1 X-Codes and X-Precoders: Encoding

For X-Codes, each symbol in \( u \) takes values from a regular \( M^2 \)-QAM constellation which consists of the \( M \)-PAM constellation \( S \overset{\Delta}{=} \{ \tau(2i - (M - 1)) \mid i = 0, 1, \ldots, (M - 1) \} \) used in quadrature on the real and the imaginary components of the channel. The
constant $\tau$ is
\[ \tau \triangleq \sqrt{\frac{3E_s}{2(M^2 - 1)}}, \quad (5.36) \]
and $E_s = \frac{P_c}{N_r}$ is the average symbol energy for each information symbol in the vector $u$. Gray mapping is used to map the bits separately to the real and imaginary component of the symbols in $u$. We fix $u^0$ to be the zero vector. In order to avoid transmitter power enhancement, we impose an orthogonality constraint on each $A_k$ and parametrize it with a single angle $\theta_k$ as
\[ A_k = \begin{bmatrix} \cos(\theta_k) & \sin(\theta_k) \\ -\sin(\theta_k) & \cos(\theta_k) \end{bmatrix}, \quad k = 1, \cdots, N_r/2. \quad (5.37) \]
We notice that 1) both $A_k$ and $G$ are orthogonal, and 2) for X-Codes we fix the angles $\theta_k$ \textit{a priori} whereas for the X-Precoders we change the angles for each channel realization.

### 5.3.2 X-Codes and X-Precoders: ML Decoding

From (5.19) and (5.20) it is obvious that two 2-D real sphere decoders (SD) are needed for each pair. Since there are $\frac{N_r}{2}$ pairs, the total decoding complexity is $N_r$ 2-D real SDs. For X-Codes the matrices $M_k$ in (5.19) and (5.20) are given by
\[ M_k = \begin{bmatrix} \lambda_{ik} \cos(\theta_k) & \lambda_{ik} \sin(\theta_k) \\ -\lambda_{jk} \sin(\theta_k) & \lambda_{jk} \cos(\theta_k) \end{bmatrix}. \quad (5.38) \]

### 5.3.3 Optimal Design of X-Codes

In order to find the best angle $\theta_k$ for the $k$th pair, we attempt to maximize $g_k(A_k)$ under the transmit power constraints. For X-Codes, let $z_k \triangleq \Re(u_k) - \Re(v_k)$ be the difference
vector between any two information vectors, which can be written as

\[ \mathbf{z}_k = \sqrt{\frac{6E_s}{(M^2 - 1)}}(p, q)^T, \quad (p, q) \in \mathbb{S}_M, \]  

(5.39)

where

\[ \mathbb{S}_M \triangleq \{(p, q) | 0 \leq |p| \leq (M - 1), 0 \leq |q| \leq (M - 1), (p, q) \neq (0, 0)\}. \]  

(5.40)

As defined in Appendix D, the generalized pairwise distance is

\[ \tilde{d}_k^2(\Re(\mathbf{u}_k), \Re(\mathbf{v}_k), \mathbf{A}_k) = 6P_T \left( p \cos(\theta_k) + q \cos(\theta_k) \right)^2 \frac{N_r}{N_r(M^2 - 1)}. \]  

(5.41)

Since \( \mathbf{A}_k \) is parameterizable with a single angle \( \theta_k \), let

\[ g_k(\theta_k, M) \triangleq g_k(A_k). \]  

(5.42)

Also, let

\[ \varphi_{p,q} \triangleq \tan^{-1} \left( \frac{q}{p} \right). \]  

(5.43)

Using (5.41) and (D.5), we have

\[ g_k(\theta_k, M) = \frac{6P_T \min_{(p, q) \in \mathbb{S}_M} (p^2 + q^2) \cos^2(\theta_k - \varphi_{p,q})}{N_r(M^2 - 1)}. \]  

(5.44)

Using (5.34), the best \( \theta_k \), denoted by \( \theta_k^* \), is given by

\[ \theta_k^* = \arg \max_{\theta_k \in [0, \frac{\pi}{4}]} \min_{(p, q) \in \mathbb{S}_M} (p^2 + q^2) \cos^2(\theta_k - \varphi_{p,q}). \]  

(5.45)
Following (5.35), the best achievable upper bound for $P'_k$ is

$$P'_k \leq (M^2 - 1) c_k \left( \frac{3 \gamma g_k(\theta^*_k, M)}{N_r (M^2 - 1)} \right)^{-\delta_k} + o(\gamma^{-\delta_k}).$$

(5.46)

**Remark 2.** It is easily shown by the symmetry of the set $S_M$ that it suffices to consider $\theta_k \in [0, \frac{\pi}{4}]$ for the maximization in (5.45). The min-max optimization problem does not have explicit analytical solutions except for small values of $M$, for example $M = 2$. But since the encoder matrices are fixed a priori, these computations can be performed off-line.

For small MIMO systems (e.g., for $2 \times 2$ MIMO), it is possible to get a tighter upper bound by evaluating the expectation in (D.3) (See Appendix D). $P'_1$ is then upper bounded as

$$P'_1 \leq \sum_{(p,q) \in S_M} \frac{(70/81)(M^2 - 1)^4 \gamma^{-4}}{M^2(p \cos(\theta_1) + q \sin(\theta_1))^6(p^2 + q^2)} + o(\gamma^{-4}),$$

(5.47)

where $\theta_1$ is the angle used for the only pair. For larger MIMO systems, it is preferable to use the inequality in (D.4) (See Appendix D), since evaluating the expectation containing two singular values is tedious. In Fig. 5.1, we compare the WEP of a $2 \times 2$
MIMO system with the upper bound given by (5.47), and observe that the union bound is indeed tight at high SNR.

In Fig. 5.2, we plot the variation of the upper bound to the WEP w.r.t. the angle $\theta_1$ for the $2 \times 2$ MIMO system with 4-QAM and 16-QAM modulation. We observe that WEP is indeed sensitive to the rotation angle. With 4-QAM, the WEP worsens as the angle approaches either 0 or 45 degrees. With 16-QAM, the performance is even more sensitive to the rotation angle. Moreover, we observe that the performance is poor when the angles are chosen near 18.5, 26.6 and 33.7 degrees, corresponding to $\phi_{3,1}$, $\phi_{2,1}$, and $\phi_{3,2}$, respectively. From (D.3), it is clear that the performance at high SNR is determined by the minimum value of the distance $\|M_k(\Re(u_k) - \Re(v_k))\|^2$, which is

$$
(p^2 + q^2) \left( \lambda_{i_k}^2 \cos^2(\theta_k - \varphi_{p,q}) + \lambda_{j_k}^2 \sin^2(\theta_k - \varphi_{p,q}) \right)
$$

(5.48)

when $(p, q)$ takes values over the set $S_M$. If $\theta_k = \tan^{-1}(-p/q)$ for some $(p, q) \in S_M$, then the minimum distance is independent of $\lambda_{i_k}$ and depends only upon $\lambda_{j_k}$. This implies a loss of diversity order since the diversity order of the square fading coefficient $\lambda_{j_k}^2$
Figure 5.3: One quadrant of the set $S_M$ for $M = 2, 4$ (4-QAM, 16-QAM). The critical angles where performance degrades severely are shown to coincide with $\tan^{-1}(-p/q)$.

is less than that of $\lambda_{ik}^2$. For the case of $N_t = N_r = 2$, this would mean a reduction of diversity order from 4 to 1. The set $S_M$ and the critical angles are illustrated in Fig. 5.3.

### 5.3.4 Optimal Design of X-Precoder

For X-Precoders, the optimal rotation angle is tedious to compute due to lack of exact expressions for error probability. Just like X-Codes we resort to bounds on error performance. It is possible to get union bound expression for the error probability of the $k$th pair. However, we do not further upper bound the union bound by using (D.4), since by doing so we would have lost information about $\lambda_{jk}$. Instead, in the pairwise sum, we look for the term with the highest contribution to the union bound and try to minimize this term. The best angle for the $k$th pair is then given by

\[
\tilde{\theta}_k(\lambda_{ik}, \lambda_{jk}) = \arg \max_{\theta_k \in [0, \pi]} \min_{(p, q) \in S_M} d_k^2(p, q, \theta_k)
\]

\[
= \arg \max_{\theta_k \in [0, \pi/4]} \min_{(p, q) \in S_M} d_k^2(p, q, \theta_k),
\]  

(5.49)
where
\[
d_k^2(p, q, \theta_k) \triangleq (p^2 + q^2)(\lambda_{i_k}^2 \cos^2(\theta_k - \varphi_{p,q}) + \lambda_{j_k}^2 \sin^2(\theta_k - \varphi_{p,q})).
\] (5.50)

Just like for X-Codes, it can be shown that for the maximization in (5.49), it suffices to consider the range \([0, \frac{\pi}{4}]\) for \(\theta_k\). The optimization problem in (5.49) is difficult, but can be solved exactly for small values of \(M\). Also, the minimization over \((p, q) \in S_M\) need not be over the full set containing \(|S_M| = 4M(M - 1)\) elements. In fact, it can be shown that the number of elements to be searched is at most \((M^2 - 3M + 6)/2\). Therefore, for \(M = 4\) (16-QAM), we need not search over the full set of 48 elements, but rather it suffices to search over only 5 elements.

**Theorem 2.** For \(M = 2\) (4-QAM), the exact \(\tilde{\theta}_k(\lambda_{i_k}, \lambda_{j_k})\) is given by

\[
\begin{cases} 
\frac{\pi}{4}, & \beta_k \leq \sqrt{3} \\
\tan^{-1}[(\beta_k^2 - 1) - \sqrt{((\beta_k^2 - 1)^2 - \beta_k^2)}], & \beta_k > \sqrt{3},
\end{cases}
\] (5.51)

where
\[
\beta_k \triangleq \frac{\lambda_{i_k}}{\lambda_{j_k}}.
\] (5.52)

\(\beta_k\) is the ratio of subchannel gains, also known as the channel condition number.

**Proof:** Proof of this theorem is given in Appendix E. \(\blacksquare\)

Further, let
\[
\overline{d}^2_{k,\text{min}}(\lambda_{i_k}, \lambda_{j_k}) \triangleq \max_{\theta_k \in [0, \frac{\pi}{4}]} \min_{(p, q) \in S_M} d_k^2(p, q, \theta_k).
\] (5.53)

Using (5.53), the union bound to \(P_k^\prime\) is given by
\[
P_k^\prime \leq (M^2 - 1)\mathbb{E} \left[ Q\left(\frac{\sqrt{\overline{d}^2_{k,\text{min}}(\lambda_{i_k}, \lambda_{j_k})}}{2N_0}\right)\right].
\] (5.54)
The expectation in (5.54) is over the joint distribution of \((\lambda_{ik}, \lambda_{jk})\) and is difficult to compute analytically. We therefore use Monte-Carlo simulations to evaluate the exact error probability.

5.4 Y-Codes and Y-Precoder

5.4.1 Motivation

It is observed that the error performance at high SNR is dependent on the minimum value of the distance \(d_{k}^2(\Re(u_k), \Re(v_k), A_k)\) over all possible information vectors \(u_k \neq v_k\). Using the definition for \(d_{k}^2(\Re(u_k), \Re(v_k), A_k)\) (see Appendix D), we have

\[
d_{k}^2(\Re(u_k), \Re(v_k), A_k) = \|M_k(\Re(u_k) - \Re(v_k))\|^2 = \lambda_{ik}^2 e_{k,1}^2 + \lambda_{jk}^2 e_{k,2}^2,
\]

(5.55)

where \(e_k \triangleq A_k(\Re(u_k) - \Re(v_k))\).

Let \(\beta_k\) be the condition number of the equivalent channel for the \(k\)th pair (see Theorem 2). We have \(\beta_k \geq 1\), since \(\lambda_{ik} \geq \lambda_{jk}\). For the special case of \(\beta_k = 1\), \(d_{k}^2(\Re(u_k), \Re(v_k), A_k)\) is proportional to \(\|e_k\|^2\), which is the Euclidean distance between the code vectors. In such a scenario, it is known that for large \(M\) choosing the code vectors as points of the 2-dimensional hexagonal lattice would yield codes with good error performance. However, the design of good codes becomes difficult for values of \(\beta_k > 1\). We immediately notice that the effective euclidean distance in (5.55) gives more weight to \(e_{k,1}^2\) which is the difference of the vectors along the first component (since \(\lambda_{ik} > \lambda_{jk}\)). Since the total transmit power is constrained, codes should be designed such that the minimum separation of any two code vectors is more along the first component.

For X-Codes and X-Precoder, minimum separation was increased by rotating the QAM constellation by an optimal angle. However, with this approach, apart from gaining separation along the first component, we also achieve separation along the second
Chapter 5. Large-MIMO Precoding Using X- and Y-Codes

It is noted that the same diversity order can be achieved even if the minimum separation along the second component is small. Since the average transmit power is constrained, optimal code design would try to choose code vectors such that for the same transmit power, more separation is achieved along the first component (without caring much about the separation achieved along the second component). This observation along with the motivation of further reducing the decoding complexity leads to the design of Y-Codes and Y-Precoder.

### 5.4.2 Y-Codes and Y-Precoders: Encoding

The matrices $A_k$ have the structure

$$A_k = \begin{bmatrix} a_k & 2a_k \\ 2b_k & 0 \end{bmatrix}, \quad (5.56)$$

where $a_k, b_k \in \mathbb{R}^+$. Let $S_k$ be the set of pairs of integers defined by the Cartesian product

$$S_k \triangleq \left\{ [0,1] \times \left[ 0, \ldots, \frac{M}{2} - 1 \right] \right\}. \quad (5.57)$$

For e.g., with $M = 4$, the set $S_k$ is given by

$$S_k = \{ [0,0]^T, [0,1]^T, [1,0]^T, [1,1]^T \}. \quad (5.58)$$

We consider the 2-D codebook of cardinality $M$ generated by applying $A_k$ to the elements of $S_k$ and translating by $u_k^0$. The code vectors $Y_k(v), v = 1, \ldots, M,$ are given by

$$Y_k(v) = \begin{bmatrix} a_k \left( v - 1 - \frac{M - 1}{2} \right), \\ b_k(-1)^v \end{bmatrix}^T, \quad (5.59)$$

The real and imaginary components of the displacement vector for the $k$th pair, $u_k^0,$ are
Chapter 5. Large-MIMO Precoding Using X- and Y-Codes

given by

\[
\Re(u_0^k) = \Im(u_0^k) = \left[ -\frac{(M-1)a_k}{2}, -b_k \right]^T. \tag{5.60}
\]

Due to the transmit power constraint in (5.14), \(a_k\) and \(b_k\) must satisfy

\[
b_k^2 + a_k^2 \frac{M^2 - 1}{12} = \frac{P_T}{N_r}. \tag{5.61}
\]

Information bits are Gray mapped to codebook indices in such a way that the Hamming distance between bit vectors corresponding to close by (in terms of Euclidean distance) code vectors is as small as possible. The only difference between Y-Codes and Y-Precoders is that, for Y-Codes the parameters \(a_k\) and \(b_k\) are fixed \textit{a priori}, whereas for the Y-Precoders these are chosen every time the channel changes.

### 5.4.3 Y-Codes and Y-Precoders: ML Decoding

Using our codebook notation, the ML decoding rule in (5.19) and (5.20) can be equivalently written as

\[
\hat{v}_k^{(I)} = \arg \min_{v \in \{0, \ldots, (M-1)\}} \|\Re(r_k) - \Lambda_k Y_k(v)\|^2, \tag{5.62}
\]

\[
\hat{v}_k^{(Q)} = \arg \min_{v \in \{0, \ldots, (M-1)\}} \|\Im(r_k) - \Lambda_k Y_k(v)\|^2, \tag{5.63}
\]

where \(\hat{v}_k^{(I)}\) and \(\hat{v}_k^{(Q)}\) are ML estimates of the codeword indices transmitted on the real and imaginary components for the \(k\)th pair.

We next discuss a low complexity algorithm for the optimization problem in (5.62). The algorithm is the same for all pairs, and the same for both the real and imaginary components of each pair. Therefore, we only discuss the algorithm for the real component. We first partition the 2-D received signal space \((\mathbb{R}^2)\) into \(\left(\frac{M}{2} + 1\right)\) regions as
Chapter 5. Large-MIMO Precoding Using X- and Y-Codes

follows.

\[ R_0 : \left\{ [x, y]^T \in \mathbb{R}^2 \mid -\infty \leq \left( \frac{x}{\lambda_i a_k} + \frac{M - 1}{2} \right) \leq 1 \right\} , \quad (5.64) \]

\[ R_M : \left\{ [x, y]^T \in \mathbb{R}^2 \mid (M - 1) \leq \left( \frac{x}{\lambda_i a_k} + \frac{M - 1}{2} \right) \leq +\infty \right\} , \quad (5.65) \]

\[ R_i : \left\{ [x, y]^T \in \mathbb{R}^2 \mid (2i - 1) \leq \left( \frac{x}{\lambda_i a_k} + \frac{M - 1}{2} \right) \leq (2i + 1) \right\} , \quad (5.66) \]

where \( i \in [1, M/2 - 1] \). In Fig. 5.4, we illustrate the 5 regions with \( M = 8 \) for the real component of the \( k \)th pair.

We next discuss a low complexity ML decoding algorithm for Y-Codes. The first step of the decoding algorithm is to find the region to which the received vector belongs. Let

\[ t_k = \left\lfloor \frac{\mathcal{R}(r_{ik}) + M + 1}{2\lambda_i a_k} \right\rfloor . \quad (5.67) \]

The received vector belongs to the region \( R_{\zeta_k} \), where \( \zeta_k \) is explicitly given by

\[ \zeta_k = \begin{cases} 
0 & t_k \leq 0 \\
\frac{M}{2} & t_k \geq \frac{M}{2} \\
t_k & \text{otherwise.} 
\end{cases} \quad (5.68) \]

For e.g., in Fig. 5.4, the received vectors \( p_1, p_2, \) and \( p_3 \) belong to \( R_0, R_1, \) and \( R_3, \) respectively. It can be shown that once the received vector is decoded to the region \( R_{\zeta_k} \), the ML code vector is one among a reduced set of at most 3 code vectors. Therefore, at most 3 Euclidean distances need to be computed to solve the ML detection problem in (5.62), as compared to computing all the \( M \) Euclidean distances in case of a brute force search. For e.g., in Fig. 5.4, for the received vector \( p_3 \in R_3, \) the ML code vector is among \( Y_k(6), Y_k(7) \) or \( Y_k(8). \)

However, once we know the region of the received vector, it is possible to directly find
Figure 5.4: Received signal space for the real component of the \( k \)th pair. \( M = 8 \) and so we have 5 regions with vertical dashed lines demarcating the boundary between the regions. The scaled codebook vectors are represented by small filled circles along with their corresponding codebook index number. Dotted lines demarcate the boundary between the ML decision regions.

the ML code vector even without computing the 3 Euclidean distances. This involves just checking a few linear relations between the 2 components of the received vector. Therefore, the ML decoding complexity of Y-Codes is the same as that of a scalar channel. For e.g., in Fig. 5.4, the received vector \( p_3 \) is to the right of the perpendicular bisector between \( Y_k(6) \) and \( Y_k(8) \). \( p_3 \) is also above the perpendicular bisector between \( Y_k(7) \) and \( Y_k(8) \). From these two checks it can be easily concluded that the ML code vector is \( Y_k(8) \). Due to the structure of the codebook, the ML decision regions can be very easily outlined. In Fig. 5.4, the dotted lines demarcate the boundary of the ML decision regions. The hatched region illustrates the ML decision region of \( Y_k(5) \).

5.4.4 Optimal Design of Y-Codes

Given the optimal pairing in (5.32), the next step towards designing optimal Y-Codes is to find the optimal value of \((a_k, b_k)\) which minimizes the average error probability. For Y-Codes, once chosen, \((a_k, b_k)\) are fixed and do not change with every channel
realization. Since the ML decision regions are known precisely, it is possible to calculate the exact error probability. With our new codebook notation, we identify code vectors by their index in the codebook, and the error probability is given by

\[ P'_k = \frac{1}{M} \sum_v P'_k(v), \quad (5.69) \]

where \( P'_k(v) \) is the probability of error when the code vector \( Y_k(v) \) is transmitted. \( P'_k(v) \) is given by

\[
P'_k(v) = \begin{cases} 
\mathbb{E}[g_1(a_k, b_k)], & 3 \leq v \leq (M - 2) \\
\mathbb{E}[g_2(a_k, b_k)], & v = 1, M \\
\mathbb{E}[g_1(a_k, b_k) - g_3(a_k, b_k)], & v = 2, (M - 1), 
\end{cases}
\]

where the expectation is over the joint distribution of \( (\lambda_{i_k}, \lambda_{j_k}) \). Let

\[
\Psi_k(x) \triangleq \frac{\sqrt{2} (2a_k \lambda_{i_k} x - a_k^2 \lambda_{i_k}^2 - 4b_k^2 \lambda_{j_k}^2)}{4b_k \lambda_{j_k} \sqrt{N_0}},
\]

and

\[
\Phi_k(x) \triangleq -\frac{\sqrt{2} (2a_k \lambda_{i_k} x + a_k^2 \lambda_{i_k}^2 + 4b_k^2 \lambda_{j_k}^2)}{4b_k \lambda_{j_k} \sqrt{N_0}}.
\]

The functions \( g_1(a_k, b_k) \), \( g_2(a_k, b_k) \) and \( g_3(a_k, b_k) \) are given by

\[
g_1(a_k, b_k) \triangleq 1 - \int_0^{\lambda_{i_k} a_k} \frac{2 e^{-\frac{x^2}{N_0}}}{\sqrt{\pi N_0}} Q(\Psi_k(x)) \, dx,
\]

\[
g_2(a_k, b_k) \triangleq 1 - \int_{-\infty}^{\lambda_{i_k} a_k} \frac{e^{-\frac{x^2}{N_0}}}{\sqrt{\pi N_0}} Q(\Psi_k(x)) \, dx,
\]

\[
g_3(a_k, b_k) \triangleq \int_{-\infty}^{-\lambda_{i_k} a_k} \frac{e^{-\frac{x^2}{N_0}}}{\sqrt{\pi N_0}} Q(\Phi_k(x)) \, dx.
\]

(5.72)
To compute the optimal \((a_k, b_k)\), we have to minimize \(P'_k\) w.r.t. \((a_k, b_k)\) subject to the transmit power constraint in (5.61). However, it is difficult to get closed-form expressions for the optimal \((a_k, b_k)\) due to the intractability of the integrals in (5.72). This difficulty is further compounded due to the evaluation of expectation over the joint distribution of \((\lambda_{ik}, \lambda_{jk})\). However, since \((a_k, b_k)\) are fixed \textit{a priori}, it is always possible to approximately compute the optimal \((a_k, b_k)\) off-line, using Monte-Carlo techniques.

### 5.4.5 Optimal Design of Y-Precoder

For the Y-Precoder, finding the optimal \((a_k, b_k)\) for each channel realization is again difficult due to the intractability of the integrals in (5.72). In the case of Y-Codes, these could be computed off-line since \((a_k, b_k)\) are fixed \textit{a priori}. However, for Y-Precoders these cannot be computed off-line, since the optimal \((a_k, b_k)\) have to be computed every time the channel changes. Therefore, we try to optimize \((a_k, b_k)\) by minimizing the union bound for \(P'_k\). The union bound is given by

\[
P'_k \leq (M - 1) \mathbb{E} \left[ Q \left( \sqrt{\frac{d_{k,\text{min}}^2(a_k, b_k) - 2N_0}{2 \lambda^2}} \right) \right],
\]

where the expectation is over the joint distribution of \((\lambda_{ik}, \lambda_{jk})\) and

\[
d_{k,\text{min}}^2(a_k, b_k) \triangleq \min_{v \neq w} \left( \lambda_{ik}^2 a_k^2 (v - w)^2 + \lambda_{jk}^2 b_k^2 ((-1)^v - (-1)^w)^2 \right),
\]

where \(v\) and \(w\) are distinct indices of the codebook.

The optimal choice of \((a_k, b_k)\), denoted by \((a^*_k, b^*_k)\), which maximizes \(d_{k,\text{min}}^2(a_k, b_k)\) for the fixed channel gain of \((\lambda_{ik}, \lambda_{jk})\), is given by the following theorem.

**Theorem 3.** The optimal value of \((a_k, b_k)\) defined as

\[
(a^*_k, b^*_k) \triangleq \arg \max_{(a_k, b_k) \in (\mathbb{R}^+)^2} \left\{ \lambda_{ik}^2 a_k^2 + a_k^2 \frac{M^2 - 1}{M^2 - 1} \frac{P_T}{N_0} \right\} d_{k,\text{min}}^2(a_k, b_k),
\]
is given by

\[
(a_k^*, b_k^*) = \begin{cases} 
\left( \sqrt{ \frac{12P_T}{N_r(M^2-1)}, 0 \right), & \beta_k^2 \geq \frac{M^2-1}{3}, \\
\left( \frac{P_r}{N_r} \frac{\beta_k}{\sqrt{\beta_k^2 + \frac{M^2-1}{3}}}, \frac{P_r}{N_r} \sqrt{\frac{1}{\beta_k^2 + \frac{M^2-1}{3}}} \right), & \beta_k^2 < \frac{M^2-1}{3},
\end{cases}
\]

(5.76)

The corresponding optimal value of \(d_{k,\text{min}}^2(a_k, b_k)\) is given by

\[
d_{k,\text{min}}^2(a_k^*, b_k^*) = \begin{cases} 
\frac{12P_T}{N_r(M^2-1)}\lambda_k^2, & \beta_k^2 \geq \frac{M^2-1}{3}, \\
\frac{16P_T}{N_r(M^2-1)}\lambda_k^2 \left( M^2+\frac{3}{16}\beta_k^2 \right), & \beta_k^2 < \frac{M^2-1}{3}.
\end{cases}
\]

(5.77)

**Proof:** The proof of this theorem is given in Appendix F.

If we now look back at the codebook for Y-Precoders, we notice that there is power allocation on the 2 channels through the parameters \(a_k\) and \(b_k\), which can be chosen optimally based upon the knowledge of channel gains. From (5.76), we observe that the Y-Precoders use only the first channel (the better channel) when channel condition is bad \((\beta_k^2 \geq \frac{M^2-1}{3})\). For good channel condition, power is distributed between the two channels depending on the channel condition. This adaptive nature of the Y-Precoders enables it to achieve better error performance in badly conditioned channels.

Y-Codes also have a fixed-rate allocation between the two channels of a pair, since out of the \(\log_2(M)\) bits, one bit can be used to decide whether the vector in the codebook is at even index (corresponding to the second component equal to \(+b_k\)) or at odd index (corresponding to the second component equal to \(-b_k\)). The remaining bits are then used to appropriately choose among the vectors at even or odd indices. Therefore, in a way, the proposed Y-Codes always transmit 1 bit of information on the bad channel and \(\log_2(M) - 1\) bits on the good channel. This rate allocation may not be the best, and therefore even better codebooks can be constructed. One more aspect that is important is the decoding complexity, which for the proposed scheme is low and is independent of \(M\). It would be challenging to obtain good code books with variable rate allocation and low decoding complexity. We, however, do not address this problem in this thesis.
5.5 Simulation Results

In this section, we compare the performance of X-, Y-Codes and X-, Y-Precoders with other precoders. For all the simulations, we assume $N_t = N_r$. Pairing of subchannels is given by (5.32). The optimal matrices $A_k$ are chosen as discussed previously. Comparisons are made with

i) the E-dmin (equal dmin precoder proposed in [102]),

ii) the Arithmetic mean BER precoder (ARITH-MBER) proposed in [25],

iii) the Equal Energy linear precoder (EE) based upon optimizing the minimum eigenvalue for a given transmit power constraint [26],

iv) the TH precoder based upon the idea of Tomlinson-Harashima precoding applied in the MIMO context [27]), and

v) the channel inversion (CI) known as zero-forcing precoder [103].

5.5.1 Effect of Channel Condition on Error Performance

In Fig. 5.5, we plot the error performance of all precoding schemes for a $2 \times 2$ MIMO system at $\gamma = 26$ dB, as a function of the condition number $\beta = \lambda_1/\lambda_2$. We fix the total channel gain $\lambda_1^2 + \lambda_2^2$ to 1, and the target spectral efficiency to $\eta = 8$ bps/Hz. We briefly discuss the precoding schemes which are compared to the proposed X-, Y-Codes. ARITH-MBER transmits $N_s$ symbols, each from a QAM modulation alphabet. When $N_s = 1$, 256-QAM modulation (i.e., 16-PAM on the real and imaginary component to achieve 8 bps/Hz) is used on the first component of the code vector and the second component is not used for transmission. When $N_s = 2$, 16-QAM modulation is used on both the components to get 8 bps/Hz. E-dmin is a precoding scheme in which the complex linear precoding matrix is adapted to each channel realization, but both the channels are always used (i.e., $N_s = 2$). The modulation alphabet is 16-QAM.

From Fig. 5.5, we see that schemes which are fixed and do not adapt with the varying channel have good error performance for small values of $\beta$. Performance is, however, poor with increasing $\beta$. Error performance of X-Codes is also seen to deteriorate.
Figure 5.5: Effect of the channel condition number on error performance of various precoders for a $2 \times 2$ system with target spectral efficiency equal to 8 bps/Hz.

with increasing $\beta$. The only exception are the Y-Codes and ARITH-MBER. For ARITH-MBER with $N_s = 1$, the opposite is true, since it always uses only one channel for transmission. The performance of Y-Codes is more stable with increasing $\beta$ due to the fact that the codebook is designed in such a way to maximize the minimum separation along the first component without caring much about the separation on the second component which corresponds to the weak channel.

It is also observed that both the X-, Y-Precoders appear to adapt well to the changes in the channel. However, the Y-Precoders perform better than X-Precoders for $\beta \geq 3$, and hence, for channels which are badly conditioned, Y-Precoders would have a better error performance compared to X-Precoder. We shall see later that, indeed for the Rayleigh fading channel, Y-Precoders perform better than X-Precoder. Therefore, we can see that codes which are fixed and do not change with each channel realization would have a poor error performance for large values of $\beta$, since they would waste power along the second component without any effect on the effective Euclidean distance. In fading channels, $\beta$ can be very large at times. Therefore, a good code is one which adapts to $\beta$. 
We also observe that Y-Codes and Y-Precoders have the best error performance when channel condition is bad. This justifies the fact that codes for badly conditioned channels should be designed to have more separation in the minimum distance along the component corresponding to the stronger channel.

5.5.2 Diversity Order Comparison

We next discuss the diversity order achieved by the various precoding schemes with Rayleigh fading. Let the number of subchannels used for transmission be $N_s (N_s \leq N_r)$. The diversity order achieved by the linear precoders (EE and ARITH-MBER) and THP is $(N_r - N_s + 1)(N_t - N_s + 1)$ and $(N_t - N_s + 1)$, respectively, whereas the diversity order achieved by E-dmin and X-, Y-Codes is $(N_r - \frac{N_s}{2} + 1)(N_t - \frac{N_s}{2} + 1)$. The CI scheme achieves infinite diversity, but it suffers from power enhancement at the transmitter. Among all the other schemes (except CI), we observe that E-dmin and X-, Y-Codes have the best diversity order. The subsequent simulation results assume a Rayleigh fading channel.

5.5.3 Comparison of BER Performance with Full-rate Transmission

In Fig. 5.6, we plot the BER of all precoders for $N_t = N_r = N_s = 2, 4$ and a target spectral efficiency of $2N_s$ bps/Hz. The proposed X-, Y-Precoders and E-dmin have the best error performance. The increased diversity order achieved by the pairing scheme is obvious from the higher slope of the error rate for the X-, Y-Precoders compared to a slope of order 1 for the linear precoder ARITH-MBER and THP. The performance of CI is inferior due to enhanced transmit power requirement arising from the bad conditioning of the channel. It is observed that the proposed Y-Precoders perform the best for $N_t = N_r = 2$, with E-dmin only 0.5 dB away at BER of $10^{-3}$. For $N_t = N_r = 4$, E-dmin performs better than Y-Precoders by 0.4 dB at BER of $10^{-3}$. However, E-dmin has this performance gain at a higher encoding and decoding complexity compared to
Figure 5.6: BER comparison between various precoders for $N_t = N_r = N_s = 2, 4$ and $M = 2$ (4-QAM). Target spectral efficiency is equal to $2N_s$ bps/Hz.

the Y-Precoder.

5.5.4 Comparison of BER Performance for $N_t = N_r = 2, 4$

In Fig. 5.7, we plot the BER for $N_t = N_r = 2$, and a target spectral efficiency of 4, 8 bps/Hz. It is observed that the best performance is achieved by the proposed Y-Precoder. For a target spectral efficiency of 4 bps/Hz, ARITH-MBER also has a similar performance. However, for a spectral efficiency of 8 bps/Hz, the performance of ARITH-MBER is worse than that of Y-Precoders by about 2.8 dB at a BER of $10^{-3}$. This is because, to achieve higher diversity order, linear precoders do not use all the modes of transmission (i.e., $N_s < \min(N_r, N_t)$). Hence, to achieve the same target spectral efficiency, they have to use higher order QAM, which results in loss of power efficiency.

In Fig. 5.8, we plot the BER for $N_t = N_r = 4$, and a target spectral efficiency of 8, 16 bps/Hz. For a target spectral efficiency of 8 bps/Hz, E-dmin and ARITH-MBER have the best error performance. Y-Precoders perform only about 0.5 dB away at a BER of $10^{-3}$. However, for a target spectral efficiency of 16 bps/Hz, Y-Precoders perform the best. ARITH-MBER ($N_s = 2$ with 256-QAM on both channels) performs 2.6 dB worse
Figure 5.7: BER comparison between various precoders for $N_t = N_r = 2$ and target spectral efficiency = 4, 8 bps/Hz.

Figure 5.8: BER comparison between various precoders for $N_t = N_r = 4$ and target spectral efficiency = 8, 16 bps/Hz.
than Y-Precoders at a BER of $10^{-3}$. E-dmin performs the worst and is about 3.5 dB away from Y-Precoders at a BER of $10^{-3}$. E-dmin has poor performance since the precoder proposed in [102] has been optimized for 4-QAM, and therefore it does not perform that well when the target spectral efficiency is higher than $2N_t$ bps/Hz.

### 5.5.5 Comparison of BER Performance for $N_t = N_r = 16$

In Fig. 5.9, we plot the BER for a large-MIMO system with $N_t = N_r = 16$ and perfect CSIT and CSIR. The target spectral efficiency is 64 bps/Hz. The X-, Y-Precoders and E-dmin use 16-QAM modulation symbols on all the available $\min(N_t, N_r) = 16$ modes of transmission. In order to achieve a better diversity order, the linear precoder (ARITH-MBER) uses only $N_s = 8$ symbols for transmission (note that the linear precoders achieve a diversity order of only 1, when transmitting at full rate $N_s = \min(N_t, N_r)$; see Fig. 5.6). For the ARITH-MBER precoder, since only $N_s = 8$ symbols are used, in order to meet a target spectral efficiency of 64 bps/Hz, the symbols are chosen from the 256-QAM signal set. Due to the expansion of the QAM alphabet size, ARITH-MBER has a performance inferior to X- and Y-Precoders. E-dmin has the worst BER performance, which is attributed to the fact that the precoder has been optimized only for 4-QAM modulation symbols. Y-Precoder has the best BER performance; at an un-coded BER of $10^{-3}$, it performs better than the X-Precoder, ARITH-MBER and E-dmin by about 0.6, 2.8 and 3.7 dB, respectively.

### 5.5.6 X-Codes versus Y-Codes

In Fig. 5.10, we compare the BER performance of the proposed X- and Y-Codes for a $N_t = N_r = 2$ system at spectral efficiencies of 4, 8 bps/Hz. It is observed that Y-Codes have a significant performance gain over X-Codes. For a target spectral efficiency of both 4 and 8 bps/Hz, Y-Codes perform better than X-Codes by about 1.5 dB at a BER
of $10^{-3}$. This is primarily due to the novel constellation structure of the proposed Y-Codes (as compared to the simple rotation encoder for X-Codes), which ensures that the minimal distance between constellation points does not become too small when channel is poorly conditioned.

In Fig. 5.11, we compare the BER performance of the proposed X- and Y-Codes for a $N_t = N_r = 4$ system at spectral efficiencies of 8, 16 bps/Hz. Y-Codes again perform better than X-Codes by about 0.7 dB for a spectral efficiency of 8 bps/Hz, and by about 1.5 dB for a spectral efficiency of 16 bps/Hz.

### 5.5.7 X-, Y-Codes versus X-, Y-Precoders

In this section, we discuss the performance gain achieved by optimally choosing the encoder matrices $A_k$ for each channel realization, as compared to having them fixed \textit{a priori}.

In Fig. 5.10, we compare the performance of the X-, Y-Precoders with that of X-, Y-Codes for $N_t = N_r = 2$ at target spectral efficiencies of 4, 8 bps/Hz. For a target spectral efficiency of 4 bps/Hz, X-, Y-Precoders perform only marginally better than...
Chapter 5. Large-MIMO Precoding Using X- and Y-Codes

Figure 5.10: BER comparison between the proposed X-Codes and Y-Codes for $N_t = N_r = 2$ with spectral efficiency = 4, 8 bps/Hz.

Figure 5.11: BER comparison between the proposed X-Codes and Y-Codes for $N_t = N_r = 4$ with spectral efficiency = 8, 16 bps/Hz.
X-, Y-Codes (by only about 0.2 dB at a BER of $10^{-3}$). However, for a target spectral efficiency of 8 bps/Hz, the X-Precoder performs better than X-Codes by about 1 dB, whereas Y-Precoders perform better than Y-Codes by about 0.2 dB at a BER of $10^{-3}$. Therefore, changing the encoder matrices with channel realization is beneficial for X-Codes. However, it is observed that Y-Precoders do not have as much gain in performance compared to Y-Codes. For $N_t = N_r = 4$, it is observed from Fig. 5.11 that for a target spectral efficiency of 8 bps/Hz, X-, Y-Precoders have almost similar performances as those of X-, Y-Codes. However, for a target spectral efficiency of 16 bps/Hz, X-Precoders perform better than X-Codes by about 0.7 dB, whereas Y-Precoders perform better than Y-Codes by about 0.3 dB at a BER of $10^{-3}$.

The performance gain of X-Precoders over X-Codes is much more significant as compared to that of the Y-Precoders over Y-Codes. Also, for X-Precoders, this performance gain is significant only with higher order QAM. This is due to the fact that the error performance is much more sensitive to the rotation angle for higher order QAM (see Fig. 5.2), and therefore adjusting the rotation angle with respect to the varying channel is expected to result in performance improvement.

On the other hand, Y-Precoders are only marginally better than Y-Codes irrespective of the spectral efficiency. This is attributed to the fact that for the Y-Precoders we optimize the upper bound to the probability of error rather than the exact error probability. We do this, because of the analytical intractability of the exact error probability expression. This leads to a suboptimal choice of the encoder matrices, and therefore a suboptimal error performance. This is obvious from Fig. 5.12, where we plot the exact optimal WEP in comparison with the error probability of the proposed suboptimal Y-Precoder. The exact optimal WEP (i.e., error probability with the optimal choice of encoder matrices) is computed through Monte Carlo techniques using (5.69) and the integrals in
Figure 5.12: Word error probability comparison between the proposed suboptimal Y-Precoders and exact optimal Y-Precoders for $N_t = N_r = 2, 4$ with spectral efficiency = $4N_t$ bps/Hz.

(5.72). The exact optimal error probability is better than the proposed suboptimal Y-Precoder by about 1.8 dB for a $N_t = N_r = 2$ system, and is better by about 1 dB for a $N_t = N_r = 4$ system at a WEP of $10^{-1}$. This, therefore, suggests the existence of better Y-Precoders compared to what has been proposed here.

5.6 Complexity

In this section, we discuss the computational complexity of X-, Y-Codes and compare it with those of other precoding schemes. The linear precoders (ARITH-MBER and EE), E-dmin and X-Codes need to compute the SVD decomposition of $H$. The CI and THP schemes involve computing the pseudo-inverse and QR decomposition of $H$, respectively. The complexity of computing SVD, QR as well as pseudo-inverse is $O(N_t^3)$. Since the channel is slowly fading, these computations can be performed once, and can be used until the channel changes. We, therefore, do not consider the complexity of these decompositions in the discussion below.
5.6.1 Encoding Complexity

The encoding complexity of all the schemes is \( O(N_t N_r) \), which is due to the transmit pre-processing filter. If the number of operations were to be computed, CI and X-, Y-Codes would have the lowest complexity. This is so because linear precoders need to compute an extra pre-processing matrix (in addition to SVD). THP also has to do successive interference pre-cancellation (in addition to QR). On the other hand, E-dmin and X-, Y-Codes need to only compute SVD, which automatically gives the pre-processing and the post-processing matrices. Also, X-, Y-Codes have lower encoding complexity compared to E-dmin, because the encoding matrices \( A_k \) are real, as opposed to being complex for E-dmin. CI has an even lower complexity since there is no spatial coding.

5.6.2 Decoding Complexity

The decoding complexity of all the schemes have a square dependence on \( N_r \). This is due to the post-processing matrix filter at the receiver. The linear precoders, CI and THP employ post processing at the receiver, which enables independent ML decoding for each subchannel. With QAM modulation symbols, this is only a rounding operation for each subchannel. E-dmin and X-Codes, on the other hand, use sphere decoding to jointly decode pairs of subchannels. ML decoding for X-Codes is accomplished by using \( N_r \) 2-dimensional real sphere decoders. However, E-dmin requires \( \frac{N_r}{2} \) 4-dimensional real sphere decoders. The average complexity of sphere decoding is cubic in the number of dimensions (and is invariant w.r.t modulation alphabet size \( M \)) [50], and therefore X-Codes have a much lower decoding complexity when compared to E-dmin. The ML decoding complexity of Y-Codes is independent of \( M \), and is equal to the ML decoding complexity of a scalar channel. Therefore, the linear precoders, CI, THP and Y-Codes have the lowest ML decoding complexity among the considered precoding schemes.
Finally, we remark that the good performance and low complexity of the proposed X-/Y-Codes and X-/Y-Precoders make them well suited for high spectral efficiency large-MIMO precoding. Further, we will exploit the structure of X-codes to increase MIMO capacity with discrete input alphabets in the next chapter.
Chapter 6

Precoding with X-codes to Increase Capacity with Discrete Input Alphabets

In this chapter, we propose a non-diagonal precoder based on the X-Codes proposed in the previous chapter to increase the mutual information with discrete input alphabet. Many modern communication channels are modeled as a Gaussian MIMO channel. Examples include multi-tone digital subscriber line (DSL), orthogonal frequency division multiplexing (OFDM), and multiple transmit-receive antenna systems. It is known that the capacity of the Gaussian MIMO channel is achieved by beamforming a Gaussian input alphabet along the right singular vectors of the MIMO channel. The received vector is projected along the left singular vectors, resulting in a set of parallel Gaussian subchannels. Optimal power allocation between the subchannels is achieved by waterfilling [91]. In practice, the input alphabet is not Gaussian and is generally chosen from a finite signal set.

We distinguish between two kinds of MIMO channels: i) diagonal (or parallel) channels and ii) non-diagonal channels.

For a diagonal MIMO channel with discrete input alphabets, assuming only power allocation on each subchannel (i.e., a diagonal precoder), Mercury/waterfilling was shown to be optimal by Lozano et al. in [98]. With discrete input alphabets, Cruz et al.
later proved in [95] that the optimal precoder is, however, non-diagonal, i.e., precoding needs to be performed across all the subchannels.

For a general non-diagonal Gaussian MIMO channel, it was also shown in [95] that the optimal precoder is non-diagonal. Such an optimal precoder is given by a fixed point equation, which requires a high complexity numeric evaluation. Since the precoder jointly codes all the $n$ inputs, joint decoding is also required at the receiver. Thus, the decoding complexity can be very high, specially for large $n$, as in the case of DSL, OFDM and large-MIMO systems. This motivates our quest for a practical low-complexity precoding scheme achieving near optimal capacity.

In this chapter, we consider a general MIMO channel and a non-diagonal precoder based on X-Codes proposed in the previous chapter. The MIMO channel is transformed into a set of parallel subchannels using singular value decomposition (SVD) and X-Codes are then used to pair the subchannels. X-Codes are fully characterized by the pairings and the 2-dimensional real rotation matrices for each pair. These rotation matrices are parameterized with a single angle. This precoding structure enables us to express the total mutual information as a sum of the mutual information of all the pairs.

The problem of finding the optimal precoder with the above structure, which maximizes the total mutual information, can be split into two tractable problems: i) optimizing the rotation angle and the power allocation within each pair, and ii) finding the optimal pairing and power allocation among the pairs. It is shown that the mutual information achieved with the proposed pairing scheme is very close to that achieved with the optimal precoder by Cruz et al. in [95], and is significantly better than the Mercury/waterfilling strategy by Lozano et al. in [98]. Our approach greatly simplifies both the precoder optimization and the detection complexity, making it suitable for practical applications.

The rest of this chapter is organized as follows. Section 6.1 introduces the system
model. In Section 6.2, we discuss the optimal precoder with discrete inputs in [95] and the relevant MIMO capacity. In Section 6.3, we propose a precoding scheme using X-Codes with discrete inputs and present its relevant capacity. In Section 6.4, we consider the first problem, which is to find the optimal rotation angle and power allocation for a given pair. This problem is equivalent to optimizing the mutual information for a Gaussian MIMO channel with two subchannels. In Section 6.5, using the results from Section 6.4, we attempt to optimize the mutual information for a Gaussian MIMO channel with a large number of subchannels (i.e., in a large-MIMO system). In Section 6.6, we discuss the application of our precoding to OFDM and large-MIMO systems.

6.1 System Model and Precoding with Gaussian Inputs

We consider a $N_t \times N_r$ MIMO channel, where the CSI is known perfectly at both transmitter and receiver. Let $x = (x_1, \cdots, x_{N_t})^T$ be the vector of input symbols to the channel, and let $H = \{h_{ij}\}, i = 1, \cdots, N_r, j = 1, \cdots, N_t$, be a full rank $N_r \times N_t$ channel coefficient matrix, with $h_{ij}$ representing the complex channel gain between the $j$th input symbol and the $i$th output symbol. The vector of $N_r$ channel output symbols is given by

$$y = \sqrt{P_T}Hx + w, \quad (6.1)$$

where $w$ is an uncorrelated Gaussian noise vector, such that $\mathbb{E}[ww^H] = I_{N_r}$, and $P_T$ is the total transmitted power. The power constraint is given by

$$\mathbb{E}[\|x\|^2] = 1. \quad (6.2)$$

The maximum multiplexing gain of this channel is $n = \min(N_t, N_r)$.

Let $u = (u_1, \cdots, u_n)^T \in \mathbb{C}^n$ be the vector of $n$ information symbols to be sent through the MIMO channel, with $\mathbb{E}[|u_i|^2] = 1, i = 1, \cdots, n$. Then the vector $u$ can be precoded
using a $N_t \times n$ matrix $T$, resulting in $x = Tu$.

The capacity of the deterministic Gaussian MIMO channel is then achieved by solving

**Problem 1.**

\[
C(H, P_T) = \max_{K_x | tr(K_x) = 1} I(x; y | H) \geq \max_{K_u, T | tr(TK_uT^H) = 1} I(u; y | H),
\]

where $I(x; y | H)$ is the mutual information between $x$ and $y$, and $K_x \triangleq \mathbb{E}[xx^H]$, $K_u \triangleq \mathbb{E}[uu^H]$ are the covariance matrices of $x$ and $u$, respectively. The inequality in (6.3) follows from the data processing inequality [91].

Let us consider the SVD of the channel $H = U\Lambda V$, where $U \in \mathbb{C}^{N_t \times n}$, $\Lambda \in \mathbb{C}^{n \times n}$, $V \in \mathbb{C}^{n \times N_t}$, $U^H U = V V^H = I_n$, and $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_n)$ with $\lambda_1 \geq \lambda_2, \cdots, \lambda_n \geq 0$.

Telatar showed in [5] that the Gaussian MIMO capacity $C(H, P_T)$, is achieved when $x$ is Gaussian distributed and $VK_xV^H$ is diagonal. Diagonal $VK_xV^H$ can be achieved by using the optimal precoder matrix $T = V^H P$, where $P \in (\mathbb{R}^+)^n$ is the diagonal power allocation matrix such that $\text{tr}(PP^H) = 1$. Furthermore, $u_i, i = 1, \cdots, n$, are i.i.d. Gaussian (i.e., no coding is required across the input symbols $u_i$). With this, the second line of (6.3) is actually an equality. Also, projecting the received vector $y$ along the columns of $U$ is information lossless and transforms the non-diagonal MIMO channel into an equivalent diagonal channel with $n$ non-interfering subchannels. The equivalent diagonal system model is then given by

\[
r \triangleq U^H y = \sqrt{P_T} \Delta u + \tilde{w},
\]

where $\tilde{w}$ is the equivalent noise vector, and has the same statistics as $w$. The total mutual information is now given by

\[
I(x; y | H) = \sum_{i=1}^n \log_2(1 + \lambda_i^2 p_i^2 P_T).
\]
Note that now the mutual information is a function of only the power allocation matrix $P = \text{diag}(p_1, \ldots, p_n)$, with the constraint $\text{tr}(PP^H) = 1$. Optimal power allocation is achieved through waterfilling between the $n$ parallel channels of the equivalent system in (6.4) [91].

### 6.2 Optimal Precoding with Discrete Inputs

In practice, discrete input alphabets are used. Subsequently, we assume that the $i$th information symbol is given by $u_i \in \mathcal{U}_i$, where $\mathcal{U}_i \subset \mathbb{C}$ is a finite signal set. Let $\mathcal{S} \triangleq \mathcal{U}_1 \times \mathcal{U}_2 \times \cdots \times \mathcal{U}_n$ be the overall input alphabet. The capacity of the Gaussian MIMO channel with discrete input alphabet $\mathcal{S}$ is defined by the following problem:

**Problem 2.**

$$C_S(H, P_T) = \max_{T \mid u \in \mathcal{S}, \|T\|_F^2 = 1} I(u; y|H).$$  \hspace{1cm} (6.6)

Note that there is no maximization over the pdf of $u$, since we fix $K_u = I_n$. The optimal precoder $T^{\text{opt}}$, which solves Problem 2, is given by the following fixed point equation given in [95]:

$$T^{\text{opt}} = \frac{H^HHT^{\text{opt}}E}{\|H^HHT^{\text{opt}}E\|_F},$$  \hspace{1cm} (6.7)

where $E$ is the minimum mean-square error (MMSE) matrix of $u$ given by

$$E = \mathbb{E}[(u - \mathbb{E}[u|y])(u - \mathbb{E}[u|y])^H].$$  \hspace{1cm} (6.8)

The optimal precoder is derived using the relation between MMSE and mutual information [106]. We observe that, with discrete input alphabets, it is no longer optimal to beamform along the column vectors of $V^H$ and then use waterfilling on the parallel subchannels. Even when $H$ is diagonal (parallel non-interfering subchannels), the
optimal precoder $T^{opt}$ is non-diagonal. $T^{opt}$ can be computed numerically (using a gradient based method) as discussed in [95]. However, the complexity of computing $T^{opt}$ is prohibitively high for practical applications, especially when $n$ is large and/or the channel changes frequently.

We propose a suboptimal precoding scheme based on the X-Codes proposed in the previous chapter, which achieves close to the optimal capacity $C_S(H, P_T)$, at low encoding and decoding complexities.

### 6.3 Precoding with X-Codes

X-Codes are based on a pairing of $n$ subchannels $\ell = \{(i_k, j_k) \in [1, n] \times [1, n], i_k < j_k, k = 1, \ldots, n/2\}$. For a given $n$, there are $(n - 1) \cdot (n - 3) \cdots 3 \cdot 1$ possible pairings. Let $\mathcal{L}$ denote the set of all possible pairings. For e.g., with $n = 4$, we have

$$\mathcal{L} = \{\{(1, 4), (2, 3)\}, \{(1, 2), (3, 4)\}, \{(1, 3), (2, 4)\}\}. \quad (6.9)$$

X-Codes are generated by a $n \times n$ real orthogonal matrix, denoted by $G$. When precoding with X-Codes, the precoder matrix is given by $T = V^H P G$, where

$$P = \text{diag}(p_1, p_2, \cdots, p_n) \in (\mathbb{R}^+)^n$$

is the diagonal power allocation matrix such that $\text{tr}(PP^H) = 1$. The $k$th pair consists of subchannels $i_k$ and $j_k$. For the $k$th pair, the information symbols $u_{i_k}$ and $u_{j_k}$ are jointly coded using a $2 \times 2$ real orthogonal matrix $A_k$ given by

$$A_k = \begin{bmatrix} \cos(\theta_k) & \sin(\theta_k) \\ -\sin(\theta_k) & \cos(\theta_k) \end{bmatrix}, \quad k = 1, \ldots, n/2. \quad (6.10)$$

The angle $\theta_k$ can be chosen to maximize the mutual information for the $k$th pair. Each
$A_k$ is a submatrix of the code matrix $G = (g_{i,j})$ as shown below

\begin{align*}
g_{i_k,i_k} &= \cos(\theta_k) & g_{i_k,j_k} &= \sin(\theta_k) \\
g_{j_k,i_k} &= -\sin(\theta_k) & g_{j_k,j_k} &= \cos(\theta_k).
\end{align*}

(6.11)

In the previous chapter, we showed that, for achieving the best diversity gain, an optimal pairing is one in which the $k$th subchannel is paired with the $(n - k + 1)$th subchannel. For e.g., with this pairing and $n = 6$, the X-Code generator matrix is given by

\[
G = \begin{bmatrix}
\cos(\theta_1) & \sin(\theta_1) \\
\cos(\theta_2) & \sin(\theta_2) \\
\cos(\theta_3) & \sin(\theta_3) \\
-\sin(\theta_3) & \cos(\theta_3) \\
-\sin(\theta_2) & \cos(\theta_2) \\
-\sin(\theta_1) & \cos(\theta_1)
\end{bmatrix}.
\]

(6.12)

The special case with $\theta_k = 0, k = 1, 2, \cdots, n/2$, results in no coding across subchannels.

Given the generator matrix $G$, the subchannel gains $\Lambda$, and the power allocation matrix $P$, the mutual information between $u$ and $y$ is given by

\[
I_S(u; y|\Lambda, P, G) = h(y|\Lambda, P, G) - h(w)
\]

\[
= -\int_{y \in \mathbb{C}^{N_r}} p(y|\Lambda, P, G) \log_2(p(y|\Lambda, P, G)) dy - n \log_2(\pi e),
\]

(6.13)

where the received vector pdf is given by

\[
p(y|\Lambda, P, G) = \frac{1}{|S|^n} \sum_{u \in S} e^{-\|y - \sqrt{P_T}U\Lambda P u\|^2},
\]

(6.14)

and when $n = N_r$ (i.e., $N_r \leq N_t$), it is equivalently given by

\[
p(y|\Lambda, P, G) = \frac{1}{|S|^n} \sum_{u \in S} e^{-\|y - \sqrt{P_T}\Lambda P u\|^2},
\]

(6.15)
where \( \mathbf{r} = (r_1, r_2, \ldots, r_n)^T \triangleq \mathbf{U}^H \mathbf{y} \).

We next define the capacity of the MIMO Gaussian channel when precoding with \( \mathbf{G} \).

In the following, we assume that \( N_r \leq N_t \), so that \( I_S(\mathbf{u}; \mathbf{y}|\Lambda, \mathbf{P}, \mathbf{G}) = I_S(\mathbf{u}; \mathbf{r}|\Lambda, \mathbf{P}, \mathbf{G}) \).

Note that, when \( N_r > N_t \), the receiver processing \( \mathbf{r} = \mathbf{U}^H \mathbf{y} \) becomes information lossy, and \( I_S(\mathbf{u}; \mathbf{y}|\Lambda, \mathbf{P}, \mathbf{G}) > I_S(\mathbf{u}; \mathbf{r}|\Lambda, \mathbf{P}, \mathbf{G}) \).

We introduce the following definitions. For a given pairing \( \ell \), let

\[
\mathbf{r}_k \triangleq (r_{ik}, r_{jk})^T, \quad \mathbf{u}_k \triangleq (u_{ik}, u_{jk})^T, \quad \Lambda_k \triangleq \text{diag}(\lambda_{ik}, \lambda_{jk}), \quad \mathbf{P}_k \triangleq \text{diag}(p_{ik}, p_{jk}), \quad \mathbf{S}_k \triangleq \mathbf{U}_{ik} \times \mathbf{U}_{jk}.
\]

Due to the pairing structure of \( \mathbf{G} \) the mutual information \( I_S(\mathbf{u}; \mathbf{r}|\Lambda, \mathbf{P}, \mathbf{G}) \) can be expressed as the sum of mutual information of all the \( n/2 \) pairs as follows:

\[
I_S(\mathbf{u}; \mathbf{r}|\Lambda, \mathbf{P}, \mathbf{G}) = \sum_{k=1}^{n/2} I_{S_k}(\mathbf{u}_k; \mathbf{r}_k|\Lambda_k, \mathbf{P}_k, \theta_k). \tag{6.16}
\]

Having fixed the precoder structure to \( \mathbf{T} = \mathbf{V}^H \mathbf{PG} \), we can formulate the following problem:

**Problem 3.**

\[
C_X(\mathbf{H}, P_T) = \max_{\mathbf{G}, \mathbf{P} \mid \mathbf{u} \in \mathcal{S}, \text{tr}(\mathbf{PP}^H) = 1} I_S(\mathbf{u}; \mathbf{r}|\Lambda, \mathbf{P}, \mathbf{G}). \tag{6.17}
\]

It is clear that the solution of the above problem is still a formidable task, although it is simpler than Problem 2. In fact, instead of the \( n \times n \) variables of \( \mathbf{T} \), we now deal with \( n \) variables for power allocation in \( \mathbf{P} \), \( n/2 \) variables for the angles defining \( \mathbf{A}_k \), and the pairing \( \ell \in \mathcal{L} \). In the following, we will show how to efficiently solve Problem 3 by splitting it into two simpler problems.

Power allocation can be divided into power allocation among the \( n/2 \) pairs, followed by power allocation between the two subchannels of each pair.

Let \( \mathbf{ar{P}} = \text{diag}(\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_{n/2}) \) be a diagonal matrix, where \( \bar{p}_k \triangleq \sqrt{p_{ik}^2 + p_{jk}^2} \) with \( p_{ik}^2 \) being the power allocated to the \( k \)th pair. The power allocation within each pair can be simply expressed in terms of the fraction \( f_k \triangleq p_{ik}^2/\bar{p}_k^2 \) of the power assigned to the first
subchannel of the pair. The mutual information achieved by the $k$th pair is then given by

$$I_{S_k}(u_k; r_k| \Lambda_k, P_k, \theta_k) = I_{S_k}(u_k; r_k| \Lambda_k, \bar{p}_k, f_k, \theta_k)$$

(6.18)

$$= - \int_{r_k \in C} p(r_k) \log_2 p(r_k) \, dr_k - 2 \log_2 (\pi e), \quad (6.19)$$

where $p(r_k)$ is given by

$$p(r_k) = \frac{1}{|S_k| \pi^2} \sum_{u_k \in S_k} e^{-\|r_k - \sqrt{P_k} A_k \bar{p}_k A_k u_k\|^2}, \quad (6.20)$$

where $F_k \triangleq \text{diag}(\sqrt{f_k}, \sqrt{1-f_k})$ and $A_k$ is given by (6.10).

The capacity of the discrete input MIMO Gaussian channel when precoding with X-Codes can be expressed as

**Problem 4.**

$$C_X(H, P_T) = \max_{\ell \in L, P|tr(PP^H)=1} \sum_{k=1}^{n/2} C_{S_k}(k, \ell, \bar{p}_k),$$

(6.21)

where $C_{S_k}(k, \ell, \bar{p}_k)$, the capacity of the $k$th pair in the pairing $\ell$, is achieved by solving

**Problem 5.**

$$C_{S_k}(k, \ell, \bar{p}_k) = \max_{\theta_k, f_k} I_{S_k}(u_k; r_k| \Lambda_k, \bar{p}_k, f_k, \theta_k).$$

(6.22)

In other words, we have split Problem 3 into two different simpler problems. Firstly, given a pairing $\ell$ and power allocation between pairs $\bar{P}$, we can solve Problem 5 for each $k = 1, 2, \ldots, n/2$. Problem 4 uses the solution to Problem 5 to find the optimal pairing $\ell^{opt}$ and the optimal power allocation $\bar{P}^{opt}$ between the $n/2$ pairs. For small $n$, the optimal pairing and power allocation between pairs can always be computed numerically and by brute force enumeration of all possible pairings. This is, however,
prohibitively complex for large $n$, and we shall discuss heuristic approaches in Section 6.5.

We will show in the following that, although suboptimal, precoding with X-Codes will provide a close to optimal capacity with the additional benefit that the detection complexity at the receiver is highly reduced, since there is coupling only between pairs of channels, as compared to the case of full-coupling for the optimal precoder in [95].

In the next section, we solve Problem 5, which is equivalent to finding the optimal rotation angle and power allocation for a Gaussian MIMO channel with only $n = 2$ subchannels.

### 6.4 Gaussian MIMO Channels with $n = 2$

With $n = 2$, there is only one pair and only one possible pairing. Therefore, we drop the subscript $k$ in Problem 5 and find $C_X(H, P_T)$ in Problem 3. The processed received vector $r \in \mathbb{C}^2$ is given by

$$r = \sqrt{P_T}F \Lambda F u + z$$

where $z = U^H w$ is the equivalent noise vector with the same statistics as $w$. Let $\alpha \overset{\Delta}{=} \lambda_1^2 + \lambda_2^2$ be the overall channel power gain and $\beta \overset{\Delta}{=} \lambda_1 / \lambda_2$ be the condition number of the channel. Then (6.23) can be re-written as

$$r = \sqrt{\tilde{P}_T} \tilde{\Lambda} F u + z,$$

where $\tilde{P}_T \overset{\Delta}{=} P_T \alpha$ and $\tilde{\Lambda} \overset{\Delta}{=} \Lambda / \sqrt{\alpha} = \text{diag}(\beta / \sqrt{1 + \beta^2}, 1 / \sqrt{1 + \beta^2})$. The equivalent channel $\tilde{\Lambda}$ now has a gain of 1, and its channel gains are dependent only upon $\beta$. Our goal is, therefore, to find the optimal rotation angle $\theta^{opt}$ and the fractional power allocation $f^{opt}$, which maximize the mutual information of the equivalent channel with condition number $\beta$ and gain $\alpha = 1$. The total available transmit power is now $\tilde{P}_T$. 

It is difficult to get analytic expressions for the optimal $\theta_{\text{opt}}$ and $f_{\text{opt}}$, and therefore we can use numerical techniques to evaluate them and store them in look-up tables to be used at run time. For a given application scenario, given the distribution of $\beta$, we decide upon a few discrete values of $\beta$ which are representative of the actual values observed in real channels. For each such quantized value of $\beta$, we numerically compute a table of the optimal values $f_{\text{opt}}$ and $\theta_{\text{opt}}$ as a function of $\tilde{P}_T$. These tables are constructed off-line. During the process of communication, the transmitter knows the value of $\alpha$ and $\beta$ from channel measurements. It then finds the look-up table with the closest value of $\beta$ to the measured one. The optimal values $f_{\text{opt}}$ and $\theta_{\text{opt}}$ are then found by indexing the appropriate entry in the table with $\tilde{P}_T$ equal to $P_T \alpha$.

In Fig. 6.1, we plot the optimal power fraction $f_{\text{opt}}$ to be allocated to the stronger channel in the pair, as a function of $P_T$. The input alphabet is 16-QAM and $\beta = 1, 1.5, 2, 4, 8$. For $\beta = 1$, both channels have equal gains, and therefore, as expected, the optimal power allocation is to divide power equally between the two subchannels. However, with increasing $\beta$, the power allocation becomes more asymmetrical. It is observed that at low $P_T$, it is optimal to allocate all power to the stronger channel. At high $P_T$, the opposite is true, and it is the weaker channel which gets most of the power. For a fixed $\beta$, as $P_T$ increases, the power allocated to the stronger channel is shifted to the weaker channel. For a fixed $P_T$, a higher fraction of the total power is allocated to the weaker channel with increasing $\beta$. In the high $P_T$ regime, these results are in contrast with the waterfilling scheme, where almost all subchannels are allocated equal power.

In Fig. 6.2, the optimal rotation angle $\theta_{\text{opt}}$ is plotted as a function of $P_T$. The input alphabet is 16-QAM and $\beta = 1, 1.5, 2, 4, 8$. For $\beta = 1$, it is observed that the mutual information is independent of $\theta$ for all values of $P_T$. For $\beta = 1.5, 2$, the optimal rotation angle is almost invariant to $P_T$. For larger $\beta$, the optimal rotation angle varies with $P_T$ and approximately ranges between $30 - 40^\circ$ for all $P_T$ values of interest.

Figure 6.3 shows the variation of the mutual information with the power fraction $f$ for
Chapter 6. Precoding with X-codes to Increase Capacity with Discrete Input Alphabets

Figure 6.1: Plot of $f^{\text{opt}}$ versus $P_T$ for $n = 2$ parallel channels with $\beta = 1, 1.5, 2, 4, 8$ and $\alpha = 1$. Input alphabet is 16-QAM.

Figure 6.2: Plot of $\theta^{\text{opt}}$ versus $P_T$ for $n = 2$ parallel channels with $\beta = 1.5, 2, 4, 8$ and $\alpha = 1$. Input alphabet is 16-QAM.
Chapter 6. Precoding with X-codes to Increase Capacity with Discrete Input Alphabets

Figure 6.3: Mutual Information of X-Codes versus power allocation fraction $f$ for $n = 2$ parallel channels with $\beta = 1, 1.5, 2, 4, 8, \alpha = 1$ and $P_T = 17$ dB. Input alphabet is 16-QAM.

$\alpha = 1$. The power $P_T$ is fixed at 17 dB and the input alphabet is 16-QAM. For a given power fraction $f$, the mutual information is maximized w.r.t. the rotation angle $\theta$. We observe that for all values of $\beta$, the mutual information is a concave function of $f$. We also observe that the sensitivity of the mutual information to variation in $f$ increases with increasing $\beta$. However, for all $\beta$, the mutual information is fairly stable (has a ‘plateau’) around the optimal power fraction. This is good for practical implementation, since this implies that an error in choosing the correct power allocation would result in a very small loss in the achieved mutual information.

In Fig. 6.4, we plot the variation of the mutual information w.r.t. the rotation angle $\theta$. The power $P_T$ is fixed at 17 dB and the input alphabet is 16-QAM. For a given rotation angle $\theta$, the mutual information is maximized w.r.t. the power allocation fraction $f$. For $\beta = 1$, the mutual information is obviously constant with $\theta$. With increasing $\beta$, mutual information is observed to be increasingly sensitive to $\theta$. However, when compared with Fig. 6.3, it can also be seen that the mutual information appears to be more sensitive to the power allocation fraction $f$ than to $\theta$. 
Chapter 6. Precoding with X-codes to Increase Capacity with Discrete Input Alphabets

Figure 6.4: Mutual information of X-Codes versus rotation angle $\theta$ for $n = 2$ parallel channels with $\beta = 1, 1.5, 2, 4, 8$, $\alpha = 1$ and $P_T = 17$ dB. Input alphabet is 16-QAM.

Figure 6.5: Mutual information versus $P_T$ for X-Codes for different $\theta$'s, $n = 2$ parallel channels, $\alpha = 1$, $\beta = 2$, and 4-QAM input alphabet.
In Fig. 6.5, we plot the mutual information of X-Codes for different rotation angles with $\alpha = 1$ and $\beta = 2$. For each rotation angle, the power allocation is optimized numerically. We observe that the mutual information is quite sensitive to the rotation angle except in the range 30-40°.

We next present some simulation results to show that indeed our simple precoding scheme can significantly increase the mutual information, compared to the case of no precoding across subchannels (i.e., Mercury/waterfilling). For the sake of comparison, we also present the mutual information achieved by the waterfilling scheme with discrete input alphabets.

We restrict the discrete input alphabets $U_i, i = 1, 2$, to be square $M$-QAM alphabets consisting of two $\sqrt{M}$-PAM alphabets in quadrature. Mutual information is evaluated by solving Problem 5 (i.e., numerically maximizing w.r.t. the rotation angle and power allocation).

In Fig. 6.6, we plot the maximal mutual information versus $P_T$, for a system with two subchannels, $\beta = 2$ and $\alpha = 1$. Mutual information is plotted for 4- and 16-QAM signal sets. It is observed that for a given achievable mutual information, coding
across subchannels is more power efficient. For e.g., with 4-QAM and an achievable mutual information of 3 bits, X-Codes require only 0.8 dB more transmit power when compared to the ideal Gaussian signaling with waterfilling. This gap increases to 1.9 dB for Mercury/waterfilling and 2.8 dB for the waterfilling scheme with 4-QAM as the input alphabet. A similar trend is observed with 16-QAM as the input alphabet.

The proposed precoder clearly performs better than Mercury/waterfilling, since the mutual information is optimized w.r.t. the rotation angle $\theta$ and power allocation, while Mercury/waterfilling, as a special case of X-Code, only optimizes power allocation and fixes $\theta = 0$.

In Fig. 6.7, we compare the mutual information achieved by X-Codes and the Mercury/waterfilling strategy for $\alpha = 1$ and $\beta = 1, 2, 4$. The input alphabet is 4-QAM. It is observed that both the schemes have the same mutual information when $\beta = 1$. 

Figure 6.7: Mutual information versus $P_T$ for $n = 2$ parallel channels with varying $\beta = 1, 2, 4$, $\alpha = 1$ and 4-QAM input alphabet.
However, with increasing $\beta$, the mutual information of Mercury/waterfilling strategy is observed to degrade significantly at high $P_T$, whereas the performance of X-Codes does not vary as much. The degradation of mutual information for the Mercury/waterfilling strategy is explained as follows. For the Mercury/waterfilling strategy, with increasing $\beta$, all the available power is allocated to the stronger channel till a certain transmit power threshold. However, since finite signal sets are used, mutual information is bounded from above until the transmit power exceeds this threshold. This also explains the reason for the intermediate change of slope in the mutual information curve with $\beta = 4$ (see the rightmost curve in Fig. 6.7). On the other hand, due to coding across subchannels, this problem does not arise when precoding with X-Codes. Therefore, in terms of achievable mutual information, rotation coding is observed to be more robust to ill-conditioned channels.

For low values of $P_T$, mutual information of both the schemes are similar, and improves with increasing $\beta$. This is due to the fact that, at low $P_T$, mutual information increases linearly with $P_T$, and therefore all power is assigned to the stronger channel. With increasing $\beta$, the stronger channel has an increasing fraction of the total channel gain, which results in increased mutual information.

In Fig. 6.8, the mutual information with X-Codes is plotted for $\beta = 1, 1.5, 2, 4, 8$ and with 16-QAM as the input alphabet. It is observed that at low values of $P_T$, a higher value of $\beta$ is favorable. However, at high $P_T$, with 16-QAM input alphabets, the performance degrades with increasing $\beta$. This degradation is more significant compared to the degradation observed with 4-QAM input alphabets. Therefore, it can be concluded that the mutual information is more sensitive to $\beta$ with 16-QAM input alphabets as compared to 4-QAM.
Chapter 6. Precoding with X-codes to Increase Capacity with Discrete Input Alphabets

Figure 6.8: Mutual information with X-Codes versus $P_T$ for $n = 2$ parallel channels with varying $\beta = 1, 1.5, 2, 4, 8, \alpha = 1$ and 16-QAM input alphabet.

6.5 Gaussian MIMO Channels with $n > 2$

We now consider the problem of finding the optimal pairing and power allocation between pairs for different Gaussian MIMO channels with even $n$ and $n > 2$. We first observe that mutual information is indeed sensitive to the chosen pairing, and this therefore justifies the criticality of computing the optimal pairing. This is illustrated through Fig. 6.9 for $n = 4$ with a diagonal channel $\Lambda = \text{diag}(0.8, 0.4, 0.4, 0.2)$ and 16-QAM. Optimal power allocation between the two pairs is computed numerically. It is observed that the pairing $\{(1, 4), (2, 3)\}$ performs significantly better than the pairing $\{(1, 3), (2, 4)\}$.

In Fig. 6.10, we compare the mutual information achieved with optimal precoding [95] to that achieved by the proposed precoder with 4-QAM input alphabet. The $4 \times 4$ full channel matrix (non-diagonal channel) is given by (42) in [95]. For X-Codes, the optimal pairing is $\{(1, 4), (2, 3)\}$ and the optimal power allocation between the pairs is computed numerically. It is observed that X-Codes perform very close to the optimal
Figure 6.9: Mutual information versus $P_T$ with two different pairings for a $n = 4$ diagonal channel and 16-QAM input alphabet.

Figure 6.10: Mutual information versus $P_T$ for the Gigabit DSL channel given by (42) in [95].
precoding scheme. Specifically, for an achievable mutual information of 6 bits, compared to the optimal precoder [95], X-Codes need only 0.4 dB extra power, whereas 2.3 dB extra power is required with Mercury/waterfilling.

Another application is in wireless MIMO channels with perfect CSI at both the transmitter and the receiver. The channel coefficients are modeled as i.i.d complex normal random variables with unit variance.

In Fig. 6.11, we plot the ergodic capacity (i.e., the mutual information averaged over channel realizations) for a $4 \times 4$ wireless MIMO channel with 16-QAM as the input alphabet. For X-Codes, the best pairing and power allocation between pairs are chosen numerically using the optimal $\theta$ and power fraction tables created off-line. It is observed that at high $P_T$, simple rotation based coding using X-Codes improves the mutual information significantly, when compared to Mercury/waterfilling. For e.g., for a target mutual information of 12 bits, X-Codes perform 1.2 dB away from the idealistic Gaussian signaling scheme. This gap from the Gaussian signaling scheme increases to 3.1 dB for the Mercury/waterfilling scheme and to 4.4 dB for the waterfilling scheme with 16-QAM alphabet.
In this application scenario, where the channel is varying, the low complexity of our precoding scheme becomes an essential feature, since the precoder can be computed on the fly using the look-up tables for each channel realization.

### 6.6 Application to OFDM and Large-MIMO

In OFDM and Large-MIMO applications, $n$ is large and Problem 4 becomes too complex to solve, since we can no more find the optimal pairing by enumeration.

It was observed in Section 6.4, that for $n = 2$, a larger value of the condition number $\beta$ leads to a higher mutual information at low values of $P_T$ (low SNR). Therefore, we conjecture that pairing the $k$th subchannel with the $(n/2 + k)$th subchannel could have mutual information very close to optimal, since this pairing scheme attempts to maximize the minimum $\beta$ among all pairs. We shall call this scheme the ‘conjectured’ pairing scheme, and the X-Code scheme, which pairs the $k$th with the $(n - k + 1)$th subchannel, the ‘X-pairing’ scheme. Note that the ‘X-pairing’ scheme was proposed in the previous chapter as a scheme which achieved the optimal diversity gain when precoding with X-Codes.

Given a pairing of subchannels, it is also difficult to compute the optimal power allocation between pairs $P$. However, it was observed that for channels with large $n$, even waterfilling power allocation between the pairs (with $\alpha_k \triangleq \sqrt{\lambda_{ik}^2 + \lambda_{jk}^2}$ as the channel gain of the $k$th pair) results in good performance.

Apart from the ‘conjectured’ and the ‘X-pairing’ schemes, we propose the following scheme which is based on the ‘Hungarian’ assignment algorithm [107] and which attempts to find a good approximation to the optimal pairing. We shall call this as the ‘Hungarian’ pairing scheme. Before describing the ‘Hungarian’ pairing scheme, we briefly review the Hungarian assignment problem as follows.

**Hungarian Assignment Problem:** Consider $m$ different workers and $m$ different jobs that have to be completed. Also, let $C(i, j)$ be the cost involved when the $i$th worker is
assigned to the $j$th job. We can therefore think of a cost matrix, whose $(i, j)$th entry has the value $C(i, j)$. The Hungarian assignment problem, is to then find the optimal assignment of workers to jobs (each worker getting assigned to exactly one job) such that the total cost of getting all the jobs completed is minimized. It is easy to see, that a maximization job assignment problem could be posed into a minimization problem and vice versa.

To find a good approximation to the optimal pairing, we split the $n$ subchannels into two groups i) Group-I: subchannels 1 to $n/2$, with the $j$th subchannel in the role of the $j$th job ($j = 1, 2, \ldots, n/2$), ii) Group-II: subchannels $n/2 + 1$ to $n$, with the $(n/2 + i)$th subchannel in the role of the $i$th worker ($i = 1, 2, \ldots, n/2$). Therefore, there are $n/2$ workers and jobs. For a given SNR $P_T$, we initially assume uniform power allocation between all pairs, and therefore assign a power of $2P_T/n$ to each pair. The value of $C(i, j)$ is evaluated by finding the optimal mutual information achieved by an equivalent $n = 2$ channel with the $(n/2 + i)$th and the $j$th subchannels as its two subchannels. This can be obtained by first choosing a table (see Section 6.4) with the closest value of $\beta$ to the given $\lambda_j/\lambda_{n/2+i}$, and then indexing the appropriate entry into the table with SNR $= 2P_T(\lambda_j^2 + \lambda_{n/2+i}^2)/n$. The Hungarian algorithm then finds the pairing with the highest mutual information. Power allocation between the pairs is then achieved through the waterfilling scheme.

It was observed through Monte-Carlo simulations that, even uniform power allocation between the subchannels results in almost same mutual information as achieved through waterfilling between pairs. This can be explained from the fact that by separating into a group of stronger (Group-I) and a group of weaker channels (Group-II), any pairing would result in all pairs having almost the same channel gain $\alpha_k$. This therefore implies that the optimal power allocation scheme would allocate nearly equal power to all pairs, which both the uniform and the waterfilling schemes would also do. Henceforth, it can be conjectured that with the proposed separation of subchannels into 2
groups, both the uniform and the waterfilling power allocation schemes would have close to optimal performance, and any further improvement in mutual information by optimizing the power allocation would be minimal. This also supports the initial usage of uniform power $2P_T/n$ to compute the entries $C(i, j)$ before executing the Hungarian algorithm. Furthermore, the computational complexity of the Hungarian algorithm is $O(n^3)$ and is therefore practically feasible.

To study the sensitivity of the mutual information to the pairing of subchannels, we also consider a ‘Random’ pairing scheme. In the ‘Random’ pairing scheme, we first choose a large number ($\approx 50$) of random pairings. For each chosen random pairing, we evaluate the mutual information (through Monte-Carlo simulations) with waterfilling power allocation between pairs. Finally, the average mutual information is computed. This gives us insight into the mean value of the mutual information w.r.t. pairing. It would also help us in quantifying the effectiveness of the heuristic pairing schemes discussed above.

**Performance in a 32-subcarrier OFDM system:** We next illustrate the mutual information
Chapter 6. Precoding with X-codes to Increase Capacity with Discrete Input Alphabets

Figure 6.13: Mutual information versus per subcarrier SNR for a OFDM system with 32 carriers. Comparison of heuristic pairing schemes.

achieved by these heuristic schemes for a OFDM system with \( n = 32 \) subchannels (i.e., subcarriers) and 16-QAM. The channel impulse response is \([-0.454 + j0.145, -0.258 + j0.198, 0.0783 + j0.069, -0.408 - j0.396, -0.532 - j0.224]\). For the ‘conjectured’ and the ‘X-pairing’ schemes also, power allocation is achieved through waterfilling between the pairs.

In Fig. 6.12 the total mutual information is plotted as a function of the SNR per subcarrier. It is observed that the proposed precoding scheme performs much better than the Mercury/waterfilling scheme. The proposed precoder with the ‘Hungarian’ pairing scheme performs within 1.1 dB of the Gaussian signaling scheme for an achievable total mutual information of 96 bits (i.e., a rate of \( 96/128 = 3/4 \)). The proposed precoder with the ‘Hungarian’ pairing scheme performs about 1.6 dB better than the Mercury/waterfilling scheme. The ‘X-pairing’ scheme performs better than the Mercury/waterfilling and worse than the ‘Hungarian’ pairing scheme. Even at a low rate of \( 1/2 \) (i.e., a total mutual information of 64 bits), the proposed precoder with the ‘Hungarian’ pairing scheme performs about 0.7 dB better than the Mercury/waterfilling
scheme.

In Fig. 6.13, we compare the mutual information achieved by the various heuristic pairing schemes. It is observed that the ‘conjectured’ pairing scheme performs very close to the ‘Hungarian’ pairing scheme except at very high SNR. For e.g., even for a high mutual information of 96 bits, the ‘Hungarian’ pairing scheme performs better than the ‘conjectured’ pairing scheme by only about 0.2 dB. However, at very high rates (like 7/8 and above), the ‘Hungarian’ pairing scheme is observed to perform better than the ‘conjectured’ pairing scheme by about 0.7 dB. Therefore, for low to medium rates, it would be better to use the ‘conjectured’ pairing since it has the same performance at a lower computational complexity. The mutual information achieved by the ‘Random’ pairing scheme is observed to be strictly inferior than the ‘conjectured’ pairing scheme at all values of SNR, and at low SNR it is even worse than the Mercury/waterfilling strategy. This, therefore, implies that the total mutual information is indeed sensitive to the chosen pairing. Further, till a rate of 1/2 (i.e., a mutual information of 64 bits) it appears that any extra optimization effort would not result in significant performance improvement for the ‘conjectured’ pairing scheme, since it is already very close to the idealistic Gaussian signaling schemes. However, at higher rate and high SNR, it may still be possible to improve the mutual information by further optimizing the selection of pairing scheme and power allocation between pairs. This is, however, a difficult problem that requires further investigation.

Performance in a $16 \times 16$ Wireless MIMO System: In Fig. 6.14, the ergodic capacity of a Rayleigh faded $16 \times 16$ wireless MIMO system is plotted as a function of $P_T$. The input alphabet used is 16-QAM. Using the Hungarian pairing scheme, X-Codes perform very close to the optimal Gaussian signaling scheme. Even at a rate of 0.5 (i.e., 32 bps/Hz), X-Codes perform only about 0.2 dB away from the idealistic Gaussian signaling scheme. X-Codes also perform significantly better than the Mercury/waterfilling and the Waterfilling schemes. At a rate of 0.75 (i.e., 48 bps/Hz), Mercury/waterfilling
Figure 6.14: Comparison of $16 \times 16$ wireless MIMO ergodic capacity with perfect CSIT/CSIR for 16-QAM input alphabet.

and the Waterfilling schemes perform about 2.0 and 4.5 dB away from X-Codes. It has also been observed through simulations that the ‘conjectured’ pairing scheme performs very close to the ‘Hungarian’ pairing scheme. Just as in the case OFDM, for wireless MIMO also, it is observed that suboptimal heuristic pairing schemes have significant performance advantage over the Mercury/waterfilling and the Waterfilling strategies. The heuristic schemes are observed to perform very close to the idealistic Gaussian signaling scheme even at rates as high as 0.75.
Chapter 7

NDS Precoder for Large Multiuser MISO Systems

In this chapter, we consider the problem of precoding in large multiuser MISO systems with large number of transmit antennas ($N_t$) at the base station and large number of downlink users ($N_u$), where each user has one receive antenna. Such large MISO systems are of interest because of the high capacities (sum-rates) of the order of tens to hundreds of bits/channel use possible in such systems. We propose a sub-optimal vector perturbation based low-complexity precoder, termed as norm descent search (NDS) precoder, suited for multiuser MISO systems for large $N_t, N_u$. The low complexity attribute of the precoder is achieved by searching for the perturbation vector over a reduced search space. Good BER performance coupled with its low complexity feature makes the NDS precoder suitable for large multiuser MISO systems.

This chapter is organized as follows. In Section 7.1, we present the system model. The proposed NDS precoder algorithm is presented in Section 7.2. Results and discussions are presented in Section 7.3.
Chapter 7. NDS Precoder for Large Multiuser MISO Systems

7.1 System Model

We consider a multiuser MISO system, where a BS communicates with $N_u$ users on the downlink. A block diagram of the system considered is shown in Figure 7.1. The BS employs $N_t$ transmit antennas and each downlink user is equipped with one receive antenna (i.e., $N_r = 1$). Let $\mathbf{u}_c \in \mathbb{C}^{N_u}$ be the complex information symbol vector, where the $i$th symbol in $\mathbf{u}_c$ is meant for the $i$th user, $i = 1, \cdots, N_u$. Precoding on the symbol vector $\mathbf{u}_c$ is carried out to obtain the precoded symbol vector $\mathbf{x}_c \in \mathbb{C}^{N_t}$, which is transmitted using $N_t$ transmit antennas such that the $j$th symbol of $\mathbf{x}_c$ is transmitted on the $j$th transmit antenna, $j = 1, \cdots, N_t$.

Let $y_i$ denote the received complex signal at user $i$, and $\mathbf{y}_c = [y_1 y_2 \cdots y_{N_u}]^T$. Let $\mathbf{H}_c \in \mathbb{C}^{N_u \times N_t}$ denote the channel matrix such that its $(i, j)$th entry $h_{i,j}$ is the complex channel gain from the $j$th transmit antenna to the $i$th user’s receive antenna. Assuming rich scattering, we model the entries of $\mathbf{H}_c$ as i.i.d. and $\mathcal{CN}(0,1)$. Let $\mathbf{n}_c$ denote the noise at the $i$th user, and $\mathbf{n}_c = [n_1 n_2 \cdots n_{N_u}]^T$. The elements of $\mathbf{n}_c$ are modeled as i.i.d.

The proposed precoding approach can be extended for the scenario where each user is equipped with more than one antenna.
\( \mathcal{CN}(0, \sigma^2) \). Therefore, \( y_c \) can be expressed in terms of \( H_c, x_c \), and \( n_c \) as

\[
y_c = H_c x_c + n_c. \tag{7.1}
\]

Let \( u_c, x_c, y_c, H_c, \) and \( n_c \) be decomposed into real and imaginary parts as follows:

\[
u_c = u_I + j u_Q, \quad x_c = x_I + j x_Q, \quad y_c = y_I + j y_Q, \quad H_c = H_I + j H_Q, \quad n_c = n_I + j n_Q. \tag{7.2}
\]

Further, we define \( u_r \in \mathbb{R}^{2N_u}, x_r \in \mathbb{R}^{2N_t}, H_r \in \mathbb{R}^{2N_u \times 2N_t}, y_r \in \mathbb{R}^{2N_u}, \) and \( n_r \in \mathbb{R}^{2N_u} \) as

\[
u_r = [u_I^T \quad u_Q^T]^T, \quad x_r = [x_I^T \quad x_Q^T]^T, \quad y_r = [y_I^T \quad y_Q^T]^T, \quad n_r = [n_I^T \quad n_Q^T]^T. \tag{7.3}
\]

Now, (7.1) can be written as

\[
y_r = H_r x_r + n_r. \tag{7.4}
\]

In the discussions to follow, we shall work with the real-valued system in (7.4). For notational convenience, we drop the subscripts \( r \) in (7.4) and write

\[
y = H x + n, \tag{7.5}
\]

where \( H = H_r \in \mathbb{R}^{2N_u \times 2N_t}, y = y_r \in \mathbb{R}^{2N_u}, x = x_r \in \mathbb{R}^{2N_t}, u = u_r \in \mathbb{R}^{2N_u}, \) and \( n = n_r \in \mathbb{R}^{2N_u}. \) With the above real-valued system model, the real part of the original complex information symbols (i.e., \( u_c \)) will be mapped to \([u_1, \cdots, u_{N_u}]\) and the imaginary part of these symbols will be mapped to \([u_{N_u+1}, \cdots, u_{2N_u}]\). For \( M \)-PAM modulation, \([u_{N_u+1}, \cdots, u_{2N_u}]\) will be zeros since \( M \)-PAM symbols take only real values. In the case of \( M \)-QAM, \([u_1, \cdots, u_{N_u}]\) can be viewed to be from an underlying \( M \)-PAM signal.
Chapter 7. NDS Precoder for Large Multiuser MISO Systems

7.1.1 Vector Perturbation

With the system model in the above, let \( G \in \mathbb{R}^{2N_t \times 2N_u} \) denote the precoding matrix. Therefore, the unit-norm transmitted symbol vector \( x \) can be written as

\[
x = \frac{Gu}{\|Gu\|}.
\]  

(7.6)

For e.g., for the well known ZF linear precoder with \( N_t \geq N_u \), the precoding matrix is given by

\[
G = G_{ZF} = H^T(HH^T)^{-1},
\]  

(7.7)

and the corresponding received signal vector \( y \) is given by

\[
y = \frac{u}{\|Gu\|} + n.
\]  

(7.8)

From (7.8), we see that \( \|Gu\| \) has a scaling effect on the instantaneous received SNR at the users, and for poorly conditioned channels this results in a significant loss in SNR. It is assumed that \( \|Gu\| \) is known at the receiver so that the received signal is scaled by \( \|Gu\| \) prior to detection\(^2\). Hence, in order to improve performance, \( \|Gu\| \) needs to be minimized. One technique suggested in the literature is to perturb the information symbol vector \( u \) in such a way that the perturbed vector \( \tilde{u} \) is another point in the lattice, but \( \|G\tilde{u}\| \) is much less than \( \|Gu\| \) [32]. Specifically, we can define \( \tilde{u} \) as

\[
\tilde{u} = u + \tau p,
\]  

(7.9)

\(^2\)It is observed via simulations that using \( \mathbb{E}\{\|Gu\|\} \) instead of the instantaneous value of \( \|Gu\| \) results in almost the same performance.
where \( p \in \mathbb{Z}^{2N_u} \) is the perturbation vector and \( \tau \) is a positive real number. The optimal value of \( \tilde{u} \), denoted by \( \tilde{u}_{opt} \), is given by

\[
\tilde{u}_{opt} = u + \tau p_{opt},
\]

(7.10)

where

\[
p_{opt} = \arg \min_{p \in \mathbb{Z}^{2N_u}} \| G(u + \tau p) \|^2.
\]

(7.11)

Exact solution of (7.11) requires exponential complexity in \( N_u \). Approximate methods (with polynomial complexity) have been proposed in the literature to solve the problem in (7.11) [44]. Even these polynomial complexity precoders are prohibitively complex for large-MISO systems with tens to hundreds of transmit antennas/users.

Our contribution here is that we propose an approximate low-complexity solution to (7.11) that scales well for large \( N_t, N_u \); the proposed solution is given in Section 7.2.

**Detection at the Receiver:** In terms of detection at the receiver, let \( \tilde{p} \) be an approximate solution to (7.11). Then, the received signal vector (after scaling by \( \| G(u + \tau \tilde{p}) \| \)) is given by

\[
\tilde{y} = (u + \tau \tilde{p}) + \tilde{n},
\]

(7.12)

where

\[
\tilde{n} = \| G(u + \tau \tilde{p}) \| n.
\]

(7.13)

The detected symbol vector at the receiver is given by

\[
\hat{u} = \tilde{y} - \tau \left[ \frac{\tilde{y} + \frac{\tilde{y}}{2}}{\tau} \right].
\]

(7.14)
In (7.14), the operation is defined on each entry of the vector since each user gets only one entry of the vector $y$. $\tau$ is a positive real scalar whose value is fixed. Choice of the value of $\tau$ affects the overall performance. Too high a value is good as far as mitigating the effect of receiver noise is concerned (since the constellation replicas are placed far apart, and there is little probability that noise may push a point from one replica to another), but on the other hand a high value of $\tau$ results in a high value of $\|G(u+\tau \tilde{p})\|$. It has been empirically observed that a good choice of $\tau$ is given by \[ \tau = 2|c_{\text{max}}| + \delta, \] where $|c_{\text{max}}|$ is the maximum value of either the real or imaginary component of the constellation symbols, and $\delta$ is the spacing between the constellation symbols. For e.g., 16-QAM is effectively two 4-PAM constellations in quadrature (taking values of -3, -1, 1, 3 on the real and imaginary axis). Therefore, for 16-QAM, $|c_{\text{max}}|$ is 3, $\delta$ is 2, and so $\tau$ is 8. Similarly, for 4-QAM, $\tau$ is 4.

### 7.2 Proposed NDS Precoder

In this section, we present the NDS precoder (see Figure 7.2 for a block diagram), which is iterative in nature and achieves an approximate solution to the problem in (7.11). Let $\tilde{u}^{(k)}$ be the perturbed information symbol vector after the $k$th iteration. We initially start with $\tilde{u}^{(0)} = u$, where $u$ is the unperturbed information symbol vector. We perturb $\tilde{u}^{(k)}$ to get $\tilde{u}^{(k+1)}$ as

\[ \tilde{u}^{(k+1)} = \tilde{u}^{(k)} + \tau p^{(k)}, \]

where $p^{(k)} \in \mathbb{Z}^{2N_u}$ is the perturbation vector for the $(k+1)$th iteration. To reduce the computational complexity of the proposed algorithm, we constrain $p^{(k)}$ to have only
one non-zero entry\footnote{This is similar to the 1-symbol neighborhood definition in the 1-LAS algorithm for large-MIMO detection in Chapter 2.}. Let $F \triangleq G^T G$, where $G \in \mathbb{R}^{2N_t \times 2N_u}$ is the precoding matrix. Further, let $q^{(k)}$ be the power (squared-norm) of the precoded symbol vector after the $k$th iteration. Therefore, $q^{(k)}$ is given by

$$q^{(k)} = \|G \tilde{u}^{(k)}\|^2 = \tilde{u}^{(k)T} F \tilde{u}^{(k)}. \quad (7.17)$$

In the $(k+1)$th iteration, the algorithm finds a constrained integer vector $p^{(k)}$ such that $q^{(k+1)} \leq q^{(k)}$. Let

$$\Delta q^{(k+1)} \triangleq q^{(k+1)} - q^{(k)}. \quad (7.18)$$

Let $e_i$ denote a $2N_u$-dimensional vector with its $i$th entry only to be one, and all the other entries to be zero. Since we allow only one non-zero entry in $p^{(k)}$, we can express $p^{(k)}$ as a scaled integer multiple of some $e_i$, $i = 1, \ldots, 2N_u$. $\Delta q^{(k+1)}$ can be negative for more than one choice of $i$. The natural question is therefore to select the appropriate $i$. Let us denote by $\Delta q_i^{(k+1)}$, the value of $\Delta q^{(k+1)}$ when $p^{(k)}$ is a scaled integer multiple of $e_i$. For each $i$, there exists a scaling integer for $e_i$, $\lambda_i^{(k)}$, which minimizes $\Delta q_i^{(k+1)}$. Let this minimum value of $\Delta q_i^{(k+1)}$ be denoted by $\Delta q_i^{(k+1)}_{opt}$. We can therefore express $\Delta q_i^{(k+1)}_{opt}$.
as

$$\Delta q_{i,\text{opt}}^{(k+1)} = \lambda_i^{(k)} \tau^2 F_{i,i} + 2 \lambda_i^{(k)} \tau z_i^{(k)},$$  \tag{7.19}$$

where $F_{i,i}$ is the $i$th diagonal entry of $F$, $z_i^{(k)}$ is the $i$th entry of the vector

$$z^{(k)} \triangleq F \tilde{u}^{(k)},$$  \tag{7.20}$$

and

$$\lambda_i^{(k)} = \begin{align*}
&\arg\min_{\lambda \in \mathbb{Z}} \Delta q_{i}^{(k+1)} \\
&= \arg\min_{\lambda \in \mathbb{Z}} \| G(\tilde{u}^{(k)} + \lambda \tau e_i) \|^2 - \| G \tilde{u}^{(k)} \|^2 \\
&= \arg\min_{\lambda \in \mathbb{Z}} \lambda^2 F_{i,i} + 2 \lambda \tau \tilde{u}^{(k)^T} F e_i \\
&= \arg\min_{\lambda \in \mathbb{Z}} \lambda^2 F_{i,i} + 2 \lambda \tau z_i^{(k)}. 
\end{align*}$$  \tag{7.21}$$

It can be shown that the exact solution to the minimization problem in (7.21) is given by

$$\lambda_i^{(k)} = -\text{sgn}(z_i^{(k)}) \left| \frac{|z_i^{(k)}|}{\tau F_{i,i}} \right|. $$  \tag{7.22}$$

Though (7.22) gives a closed-form solution to $\lambda_i^{(k)}$, we have observed in the simulations that in cases when $\lambda_i^{(k)}$ is large, the algorithm tends to get trapped in some poor local minima early in the algorithm. In order to alleviate this phenomenon, we constrain the value of $\lambda_i^{(k)}$ to be within a set $\mathbb{S} = \{-s_{\text{max}}, -(s_{\text{max}} - 1), \ldots, (s_{\text{max}} - 1), s_{\text{max}}\}$, which is a finite subset of $\mathbb{Z}$, and $s_{\text{max}}$ denotes the maximum absolute value in $\mathbb{S}$. For e.g., for 4-QAM, we have found (through simulations) the appropriate set $\mathbb{S}$ to be $\mathbb{S} = \{-1, 0, 1\}$. If $|\lambda_i^{(k)}| > s_{\text{max}}$, then $\lambda_i^{(k)}$ is set to 0, and so is $\Delta q_{i,\text{opt}}^{(k+1)}$. If $|\lambda_i^{(k)}| \leq s_{\text{max}}$, then $\Delta q_{i,\text{opt}}^{(k+1)}$ is computed as per (7.19). We shall refer to this correction in $\lambda_i^{(k)}$ as $\lambda$-adjustment. In the $(k + 1)\text{th}$ iteration, we can therefore calculate $\Delta q_{i,\text{opt}}^{(k+1)}$ for $i = 1, \ldots, 2N_u$. Given these
values of $\lambda_i^{(k)}$, $i = 1, \cdots, 2N_w$, we update $\tilde{u}^{(k)}$ as follows:

$$\tilde{u}^{(k+1)} = \tilde{u}^{(k)} + \tau \lambda_j^{(k)} e_j,$$

(7.23)

where

$$j = \arg \min_i \Delta q_i^{(k+1)}_{i,\text{opt}}.$$  \hspace{1cm} (7.24)

The values of $\lambda_j^{(k)}$ used in (7.23) are after the $\lambda$-adjustment described above. We also need to evaluate $z^{(k+1)}$. From (7.20), we can write

$$z^{(k+1)} - z^{(k)} = F (\tilde{u}^{(k+1)} - \tilde{u}^{(k)}).$$

(7.25)

Using (7.23), we can rewrite (7.25) as

$$z^{(k+1)} = z^{(k)} + \tau \lambda_j^{(k)} f_j,$$

(7.26)

where $f_j$ refers to the $j$th column of $F$. Finally, the algorithm terminates after some iteration $n$ if

$$\min_i \Delta q_i^{(n+1)}_{i,\text{opt}} \geq 0.$$ \hspace{1cm} (7.27)

It is easy to see that the algorithm guarantees a monotonic descent in $\|G\tilde{u}^{(k)}\|^2$ with every iteration until a local minima is reached. Since i) $\lambda_i^{(k)}$ can take values only from a finite integer valued set $S$, and ii) $\|G\tilde{u}^{(k)}\|^2$ has a global minina for perturbations with $\lambda_i^{(k)} \in S$, we can see that the NDS algorithm will terminate in a finite number of iterations. The listing of the NDS algorithm is presented in Table-7.1.
1. Choose the set \( S; s_{\text{max}} = \max_{s \in S} s \)
2. \( \tilde{u}(0) = u; \quad F = G^T G; \quad k = 0 \quad (k \text{ is iteration index}) \)
3. \( z^{(0)} = F\tilde{u}^{(0)}; \tau = 2|c_{\text{max}}| + \delta \)
4. \( nsymb = 2N_u; \quad (\text{nsymb is } 2N_u \text{ for QAM and } N_u \text{ for PAM}) \)
5. for \( i = 1, 2, \cdots, nsymb \)
6. \( \lambda^{(k)}_i = -\text{sgn}(z_i^{(k)}\frac{|z_i^{(k)}|}{F_{ii}}) \)
7. if \(|\lambda^{(k)}_i| > s_{\text{max}}\) \( \lambda^{(k)}_i = 0 \)
8. \( \Delta q_{i,\text{opt}}^{(k+1)} = \lambda^{(k)}_i \tau^2 F_{i,i} + 2\lambda^{(k)}_i \tau z_i^{(k)} \)
9. end; \quad \text{(end of for in Step 5)}
10. \( \Delta q_{\text{min}} = \min_i \Delta q_{i,\text{opt}}^{(k+1)} \)
11. if \( (\Delta q_{\text{min}} \geq 0) \) goto Step 16
12. \( j = \arg \min_i \Delta q_{i,\text{opt}}^{(k+1)} \)
13. \( \tilde{u}^{(k+1)} = \tilde{u}^{(k)} + \tau \lambda^{(k)}_j e_j \)
14. \( z^{(k+1)} = z^{(k)} + \tau \lambda^{(k)}_j f_j \)
15. \( k = k + 1, \) goto Step 5
16. Terminate

### 7.2.1 Complexity of the NDS Algorithm

The complexity of the NDS algorithm is analyzed here. The per-symbol computation complexities of \( G^T G \) in Step 2 and \( z^{(0)} \) in Step 3 are \( O(N_u N_t) \) and \( O(N_u) \), respectively.

Steps 5 to 15 constitute one basic iteration of the NDS algorithm, whose per-symbol complexity is constant, i.e., \( O(1) \). The mean number of iterations till the algorithm terminates, which we have obtained through simulations, has been found to be proportional to \( N_u \) (see Figure 7.3). Putting the above individual complexities together, the overall per-symbol complexity of the NDS algorithm is \( O(N_t N_u) \). This low-complexity feature makes practical precoding for large number of users (of the order of tens to hundreds) to be feasible.
Chapter 7. NDS Precoder for Large Multiuser MISO Systems

7.3 Results and Discussions

In this section, we present the uncoded and turbo coded simulation results of the BER performance of the proposed NDS precoder. In the simulations, we consider the precoder matrix \( G \) to be either ZF or MMSE. The ZF precoding matrix \( G_{ZF} \) is given by (7.7). The MMSE precoding matrix is given by 
\[
G_{MMSE} = H^T (HH^T + \sigma^2 N_t I_{N_u})^{-1}.
\]
We will refer to the NDS precoder as the NDS-MMSE precoder when \( G_{MMSE} \) is used as the precoding matrix, and as the NDS-ZF precoder when \( G_{ZF} \) is used. We consider symmetric (\( N_u = N_t \)) as well as asymmetric (\( N_u < N_t \)) systems. We will also compare the performance of the NDS precoder with that of the vector perturbation scheme in [32] which uses sphere encoding (SE) to solve (7.11); we will refer to this scheme as VP-SE scheme.

7.3.1 NDS-MMSE versus NDS-ZF Precoder Performance

In Fig. 7.4, we compare the uncoded BER performance of the NDS-MMSE precoder with that of the NDS-ZF precoder for \( N_t = N_u = 10, 50, 200, N_r = 1, 4\text{-QAM}, \) and
perfect knowledge of the channel gains. As expected, it is observed that the NDS-MMSE precoder performs better than the NDS-ZF precoder. For $N_t = N_u = 50$ and a target BER of $10^{-3}$, NDS-MMSE requires about 5 dB less SNR when compared to NDS-ZF. Also, NDS-MMSE achieves this better performance at the same complexity order as that of NDS-ZF, making it more power efficient than NDS-ZF. It can also be observed that the NDS-MMSE precoder exhibits ‘large-system effect,’ where the BER performance for $N_t = N_u = 200$ is better than for $N_t = N_u = 50$. This is similar to the large-system effect we observed for the LAS and PDA based detection in point-to-point large-MIMO links presented in Chapters 2 and 3. The fact that it is possible to get the simulated BER performance of the NDS-MMSE precoder for such large number of users like 200 users\(^4\) illustrates the suitability of the proposed precoder for large multiuser MISO systems.

\(^4\)While tens to hundreds of single-antenna downlink users can be envisioned easily in a practical system, base stations with hundreds of antennas may look quite futuristic. However, we point to reference [100], where Thomas Marzetta observes that “…Even in short coherence intervals (say five-hundred microseconds) and low SINRs (minus-ten dB reverse, and zero dB forward) a base station comprising sixteen or more antennas can both learn the forward channel via TDD reciprocity, and transmit, with high aggregate throughput, multiple data streams to multiple single-antenna terminals. It is always
Figure 7.5: Uncoded BER performance comparison of the proposed NDS-MMSE precoder versus (i) MMSE-only precoder, and (ii) the VP-SE scheme. $N_u = 8$ and $N_t = 8, 16$, $N_r = 1$, 4-QAM.

### 7.3.2 NDS-MMSE versus MMSE-only and VP-SE Schemes

In Fig. 7.5, we compare the uncoded BER performance of the NDS-MMSE precoder with those of (i) the MMSE-only precoder (without the NDS), and (ii) the VP-SE scheme in [32], for $N_u = 8$, $N_r = 1$, and 4-QAM. Performance of these precoders for $N_t = 8$ (symmetric) and $N_t = 16$ (asymmetric) are shown. The following observations can be made from the performance plots in Fig. 7.5.

- Comparing the performances of symmetric ($N_t = N_u = 8$) and asymmetric ($N_t = 16$, $N_u = 8$) systems, we see that the asymmetric system performs significantly better. This is expected, and is because of the availability of $N_t - N_u$ additional dimensions at the transmit side for the precoders to exploit.

- Comparing the performances of the MMSE-only and the NDS-MMSE precoders, advantageous to increase the number of base station antennas. One can envision a new type of cellular structure that comprises inexpensive single-antenna terminals working with base stations having fifty or one-hundred antennas, each driven by its own tower-top amplifier of power no greater than a typical cell-phone power amplifier. In this context, we note that low-complexity near-optimal precoders like the one we propose in this chapter can fill the need for such precoding algorithms in such large multiuser MISO systems.
we see that carrying out the proposed norm-descent search prior to MMSE pre-
coding achieves much better diversity order compared to MMSE-only precoding.
Given that the additional search operation itself is of low-complexity, i.e., $O(N_u)$
per-symbol complexity, compared to the $O(N_t N_u)$ per-symbol complexity of the
$G^T G$ and MMSE operations, this improvement is quite significant.

• Comparing the performances of the proposed NDS-MMSE scheme and the VP-
SE scheme, we see that the VP-SE scheme performs better at moderate to high
SNRs. However, the NDS-MMSE performance is quite close to that of the VP-
SE scheme at these SNRs. For e.g., for $N_t = 16$, the SNR gap between VP-SE
and NDS-MMSE performances at $10^{-3}$ BER is just about 0.4 dB. The NDS-MMSE
scheme achieves such good performance at a much reduced complexity com-
pared to the exponential complexity of the VP-SE scheme in solving (7.11). Fur-
ther, perturbation based schemes perform relatively poor at low SNRs, since the
optimum choice of the perturbation vector $p$ is made based on minimizing the
average transmit power and not based on minimizing the BER. In fact, the NDS-
MMSE and MMSE-only performances are slightly better than that of the VP-SE
scheme at low SNRs.

Complexity Comparison Between NDS-MMSE and VP-SE Schemes

In Fig. 7.6, we present a complexity comparison between the NDS-MMSE and VP-SE
schemes for $N_t = N_u, N_r = 1, 4$-QAM at 15 dB SNR. We have plotted the complex-
ity, measured in terms of mean CPU run time (in seconds) needed for precoding an
information symbol vector $u$ into the transmit vector $x$. Since the complexity is depen-
dent on the channel realization $H$, we averaged it over a large number of independent
channel realizations. The measured CPU run times in the simulations are shown in the
figure. We see that the search in VP-SE scheme has exponential complexity in $N_t$, as ev-
edenced by its complexity curve running parallel to the $c_{ex} 2^{N_u}$ curve for large $N_u$. This
makes VP-SE scheme not suitable for large systems. The complexity of NDS-MMSE scheme, however, is observed to be just $O(N_u^3)$, as evidenced by the $c_3N_u^3$ line running parallel to the NDS-MMSE complexity curve for large $N_u$. Therefore, the per-symbol complexity of the proposed NDS-MMSE scheme is just $O(N_u^2)$ (since there are $N_u$ symbols per $u$ vector), making it suitable for large systems.

### 7.3.3 Nearness to Sum Capacity

Next, we present the turbo coded BER performance of the NDS-MMSE precoder and its nearness to capacity. For coded systems, a relevant metric that can be used for assessing the performance is the ergodic sum capacity of the broadcast MISO channel. The ergodic sum capacity of the model in (7.1) is given by [32]

$$C_{\text{sum}} = E \left[ \sup_{\mathbf{D} \in \mathcal{A}} \log \det \left( \mathbf{I}_{N_t} + \rho \mathbf{H}_c^H \mathbf{D} \mathbf{H}_c \right) \right], \quad (7.28)$$

where $\mathcal{A}$ is the set of $N_u \times N_u$ diagonal matrices with non-negative elements that sum to 1 (i.e., $\text{tr}(\mathbf{D}) = 1$), and $\rho$ is the average SNR defined as $1/\sigma^2$. Since there is no
closed-form expression for the optimization in (7.28), we have to evaluate it through Monte-Carlo simulations. Monte-Carlo simulations are prohibitive for large systems, and so we consider upper and lower bounds to the sum capacity in the following. We note that $D = D_{\text{CSIR}} \triangleq \frac{1}{N_t} I_{N_t}$ satisfies the trace constraint, and therefore

$$C_{\text{CSIR}} \triangleq \mathbb{E} \left[ \log \det \left( I_{N_t} + \frac{\rho}{N_t} H_c H_c^H \right) \right]$$

is a lower bound for $C_{\text{sum}}$, i.e., $C_{\text{sum}} \geq C_{\text{CSIR}}$. We also note that $C_{\text{CSIR}}$ is the ergodic capacity of a point-to-point single-user MIMO system with $N_t$ receive antennas and $N_u$ transmit antennas with CSIR only. On the other hand, receiver cooperation between the users will increase the capacity, and therefore we see that the sum capacity $C_{\text{sum}}$ is upper bounded by the capacity of a point-to-point MIMO system with $N_t$ transmit antennas and $N_u$ receive antennas with CSIT and CSIR. We shall denote this upper bound by $C_{\text{CSIT}}$.

In Fig. 7.7, we plot the upper ($C_{\text{CSIT}}$) and lower ($C_{\text{CSIR}}$) bounds to the sum capacity $C_{\text{sum}}$. It is observed that the gap between the bounds diminishes with increasing SNR, and therefore any of these bounds is a good approximation at high SNR. However, at low SNRs, there is a gap between the bounds, which diminishes as the system becomes more asymmetrical. For e.g., with $N_u = 8$ users and a target spectral efficiency of 1.5 bps/Hz for each user, the gap between the upper and lower bounds is 0.5, 0.8 and 1.3 dB for $N_t = 16, 12, \text{ and } 8$, respectively. For small systems, it has been observed through Monte-Carlo simulations that the ergodic sum capacity is almost same as the lower bound $C_{\text{CSIR}}$ [32]. We will evaluate the nearness of the NDS-MMSE scheme performance with turbo coding w.r.t to the upper bound on $C_{\text{sum}}$.

**Turbo Coded BER Performance**

Figure 7.8 shows the turbo coded BER performance of the NDS-MMSE and VP-SE schemes for $N_u = 8$, $N_t = 8, 12, 16$, 4-QAM, $N_r = 1$, and rate-3/4 turbo code. The sum
Figure 7.7: Upper and lower bounds for the ergodic sum capacity, $C_{\text{sum}}$.

rate achieved in this system is $8 \times 2 \times 3/4 = 12$ bps/Hz. The minimum SNR required to achieve a sum rate of 12 bps/Hz obtained from the upper bound on the sum capacity are also shown. From Fig. 7.8, it can be seen that the VP-SE scheme achieves vertical fall in coded BER at about 9.2, 7.8 and 7.2 dB away from the respective theoretical minimum SNRs required for $N_t = 8$, 12 and 16. It is further seen that the vertical fall for the NDS-MMSE scheme for $N_t = N_u = 8$ occurs at about 1.5 dB away from that of the VP-SE scheme. For the asymmetric cases of $N_t = 12$ and $N_t = 16$, the performance of the NDS-MMSE scheme is about 0.5 dB better than that of the VP-SE scheme.

We further note that while VP-SE scheme performance can be evaluated for small systems as shown in Fig. 7.8, evaluation of its performance for hundreds of users is prohibitively complex. On the other hand, performance in such large systems with the proposed NDS-MMSE precoder can be evaluated due its low complexity. We highlight this point in Fig. 7.9, where we have plotted the turbo coded BER performance of the NDS-MMSE scheme for a large system with $N_t = N_u = 300$, $N_r = 1$, 4-QAM, rate-3/4 turbo code, and sum rate = 450 bps/Hz. To illustrate the effect of channel estimation
errors on performance, we consider a channel estimation error model where the estimated channel matrix, $\hat{H}_c$, is taken to be $\hat{H}_c = H_c + \Delta H_c$, where $\Delta H_c$ is the estimation error matrix, the entries of which are assumed to be i.i.d complex Gaussian with zero mean and variance $\sigma_e^2$. The values of $\sigma_e^2$ considered in the simulations are 0, 0.01, 0.02. Note that $\sigma_e^2 = 0$ corresponds to the case of perfect channel estimation. The following observations can be made from Fig. 7.9.

- With perfect channel estimation (i.e., $\sigma_e^2 = 0$), the NDS-MMSE precoder achieves vertical fall in turbo coded BER at about 13 dB (i.e., about 9 to 10 dB away from the theoretical minimum SNR required). The MMSE-only precoder (without the NDS), on the other hand, achieves the vertical fall only at about 16 dB. It is noted that the order of per-symbol complexity for the NDS-MMSE and the MMSE-only schemes are the same, with NDS-MMSE scheme performing better than the MMSE-only scheme.

- The robustness of the NDS-MMSE precoder to imperfect channel estimation is superior compared to the MMSE-only precoder. For e.g., for $\sigma_e^2 = 0.02$, vertical
Figure 7.9: Turbo coded BER performance of the proposed NDS-MMSE precoder without and with channel estimation errors. $N_t = N_u = 300$, $N_r = 1$, 4-QAM, rate-3/4 turbo code, sum rate = 450 bps/Hz.

fall occurs at about 15 dB for the NDS-MMSE precoder, whereas for the MMSE-only precoder vertical fall does not occur and a high error floor results.

We note that search for other low-complexity precoding algorithms for large multiuser MISO/MIMO systems, like the proposed NDS precoder, can be a topic of further investigation. Channel estimation issues in large multiuser MISO/MIMO systems can also be investigated further.
Chapter 8

Conclusions

In this thesis, we investigated low-complexity detection and precoding algorithms that can potentially enable practical realization of large-MIMO systems with tens of antennas in wireless communication terminals.

In Chapters 2, 3, and 4, we dealt with large-MIMO detection and channel estimation. Large-MIMO precoding was the subject matter in Chapters 5 and 6. Precoding for large multiuser MISO systems was investigated in Chapter 7.

In Chapter 2, we presented a low-complexity LAS algorithm suited for detection in large V-BLAST MIMO and non-orthogonal STBC MIMO systems with tens of antennas that achieve high spectral efficiencies of the order of several tens to hundreds of bps/Hz. The algorithm was shown to exhibit large-system behavior, where the bit error performance improved with increasing number of antennas, and approached near-ML performance for large number of dimensions. We also presented a training-based iterative detection/channel estimation scheme for large-MIMO systems. Our simulation results showed that the LAS detector along with the iterative detection/channel estimation scheme achieved very good performance at low complexities. Subsequent to our reporting of the LAS algorithm for large-MIMO detection in the literature, other authors have reported FPGA implementation of $32 \times 32$ V-BLAST MIMO detection for 4-/16-/64-QAM using the proposed LAS algorithm.
In Chapter 3, we presented a performance analysis of the LAS algorithm in the large-system limit, where \( N_t, N_r \to \infty \) with \( N_t = N_r \).

In Chapter 4, we presented another low-complexity algorithm for large-MIMO detection, which is based on PDA. We showed that the PDA algorithm too exhibited large-system behavior, which, along with the low-complexity attribute, made the PDA algorithm as another promising algorithm for large-MIMO detection.

We note that with the feasibility of low-complexity high-performance detection algorithms like the proposed LAS and PDA algorithms, in conjunction with the iterative detection/channel estimation scheme, large-MIMO systems with tens of antennas at high spectral efficiencies can become practical, enabling interesting high data rate wireless applications (e.g., wireless IPTV/HDTV distribution). This can motivate the inclusion of large-MIMO architectures (e.g., \( 12 \times 12, 16 \times 16, 24 \times 24, 32 \times 32 \) MIMO systems, including those using STBCs from CDA) into wireless standards like IEEE 802.11ac and IEEE 802.16/LTE-A in their evolution to achieve high data rates (multi-gigabit rates) at increased spectral efficiencies (in excess of 50 bps/Hz).

In Chapter 5, we proposed X-, Y-Codes/Precoders for large-MIMO precoding which can achieve full-rate and high diversity at low complexity by pairing the subchannels prior to SVD precoding. It was observed that indeed pairing of channels can significantly improve the overall diversity. Among all possible pairings, pairing the \( k \)th channel with the \((N_r - k + 1)\)th subchannel was found to be optimal in terms of achieving the best diversity order. One way of pairing the subchannels is by using rotation based encoding as for X-Codes/Precoders. The proposed X-Codes/Precoders have good performance for well conditioned channels. For ill-conditioned channels, we then proposed Y-Codes/Precoders. It was shown that Y-Codes/Precoders achieve the best error performance at very low complexity, when compared to other precoders in the literature. In practice, in order to improve the overall performance, it is possible to adaptively switch between X- and Y-Codes/Precoders depending on the channel...
conditions.

In Chapter 6, we proposed a low-complexity precoding scheme based on the pairing of subchannels, which achieves near optimal capacity for Gaussian MIMO channels with discrete inputs. The low-complexity feature relates to both the evaluation of the optimal precoder matrix and the detection at the receiver. This makes the proposed scheme suitable for practical applications, even when the channels are time varying and the precoder needs to be computed for each channel realization. The simple precoder structure, inspired by the X-Codes, enabled us to split the precoder optimization problem into two simpler problems. Firstly, for a given pairing and power allocation between pairs, we need to find the optimal power fraction allocation and rotation angle for each pair. Given the solution to the first problem, the second problem is then to find the optimal pairing and the power allocation between pairs. For large number of subchannels, typical of OFDM and large-MIMO systems, we also discussed different heuristic approaches for optimizing the pairing of subchannels. The proposed precoder was shown to perform close to the optimal precoder with discrete inputs, and significantly better than the Mercury/waterfilling strategy for both diagonal and non-diagonal MIMO channels. Future work can focus on finding close to optimal pairings, and close to optimal power allocation strategies between pairs.

In Chapter 7, we presented a low-complexity NDS precoder for large multiuser MISO systems. The NDS precoder was shown to perform close to the sphere encoder based vector perturbation scheme, but at a much lower complexity. So, the NDS precoder is suited, in terms of both complexity as well as performance, for large multiuser MISO systems with tens to hundreds of downlink users. The presented precoder was also shown to be robust to channel estimation errors. The feasibility of low-complexity precoders, like the NDS precoders we presented, can potentially trigger wide interest in the theory and implementation of large multiuser MISO/MIMO systems.

In concluding this thesis, we point out that the area of low-complexity near-optimal
signal processing for large-MIMO systems is both nascent as well as hugely promising. We believe that very high spectral efficiency wireless systems employing large number of antennas will be practical in the near future, and that the work reported in this thesis can be viewed as a development in that direction.
Appendix A

Proof of Optimum Update \(l_p^{(k)}\), Chapter 2

**Theorem:** The \(l_p^{(k)}\) in (2.27) minimizes \(F(l_p^{(k)})\) in (2.25) and this minimum value is non-positive.

**Proof:** Let \(r = \left\lfloor \frac{|z_p^{(k)}|}{2a_p} \right\rfloor\). Then \(\frac{|z_p^{(k)}|}{2a_p} = r + f\), where \(0 \leq f < 1\), and so we can write

\[
\frac{|z_p^{(k)}|}{a_p} = 2r + 2f. \quad (A.1)
\]

If \(l_p^{(k)}\) were unconstrained to be any real number, then the optimal value of \(l_p^{(k)}\) is \(\frac{|z_p^{(k)}|}{a_p}\), which would lie between \(2r\) and \(2r + 2\) (as per (A.1)). Since \(F(l_p^{(k)})\) is quadratic in \(l_p^{(k)}\), it is unimodular, and hence the optimal point (with \(l_p^{(k)}\) constrained) would be either \(2r\) or \(2r + 2\). Using (2.25) and (A.1), we can evaluate \(F(2r + 2) - F(2r)\) to be

\[
F(2r + 2) - F(2r) = 4a_p(1 - 2f). \quad (A.2)
\]

Since \(a_p\) is a positive quantity, the sign of \(F(2r + 2) - F(2r)\) depends upon the sign of \((1 - 2f)\). If \(f \geq 0.5\), then \(F(2r + 2) \leq F(2r)\), and therefore \(2r + 2\) is the optimal value of \(l_p^{(k)}\). Similarly, when \(f < 0.5\), \(2r\) is the optimal value of \(l_p^{(k)}\). Therefore, it follows that indeed the rounding solution given by (2.27) is optimal. \(F(l_p^{(k)})\) is non-positive for all
values of \( l_p^{(k)} \) between zero and \( \frac{2|z_p^{(k)}|}{a_p} \). If \( f < 0.5 \), then \( 2r \) is optimal, and, from (A.1), we know that \( 2r \leq \frac{|z_p^{(k)}|}{a_p} \), and therefore \( 2r < \frac{2|z_p^{(k)}|}{a_p} \). Hence \( \mathcal{F}(2r) = \mathcal{F}^{(opt)} \) is non-positive.

Similarly, if \( f \geq 0.5 \), then \( 2r + 2 \) is optimal, and \( \mathcal{F}(2r + 2) \leq \mathcal{F}(2r) \). However, since \( 2r \) is always less than \( \frac{2|z_p^{(k)}|}{a_p} \), \( \mathcal{F}(2r) \) is non-positive and therefore \( \mathcal{F}(2r + 2) = \mathcal{F}^{(opt)} \) is non-positive.
Appendix B

Proof of Lemma 5, Chapter 3

We present the proof of Lemma 5 of Chapter 3. The proof is by mathematical induction on \(n\).

**Base Case:** For \(n = 2\), we have to show that

\[
d_p d_q \frac{\text{\(h_p^T h_q\)}}{\|h_p\|^2 + \|h_q\|^2} \xrightarrow{p} 0 \quad \text{as} \quad N_t \to \infty, \forall p, q = 1, 2, \cdots, 2N_t, \ p \neq q. \tag{B.1}
\]

We can write the random variable \(\frac{\text{\(h_p^T h_q\)}}{\|h_p\|^2 + \|h_q\|^2}\) as

\[
\frac{\text{\(h_p^T h_q/(2N_t)\)}}{\left(\|h_p\|^2 + \|h_q\|^2\right)/(2N_t)}. \tag{B.2}
\]

As \(N_t \to \infty\), by strong law of large numbers, the denominator of (B.2) converges to 1 almost surely. Also, the numerator of (B.2) can be written as

\[
\frac{\text{\(h_p^T h_q\)}}{2N_t} = \sum_{k=1}^{2N_t} h_{p,k} h_{q,k}, \tag{B.3}
\]

where \(h_{p,k}\) and \(h_{q,k}\) refer to the \(k\)th entry of the vectors \(h_p\) and \(h_q\), respectively. Each \(h_{p,k}h_{q,k}\) term in the summation in (B.3) has the same distribution and has mean 0. Therefore, by strong law of large numbers, we can see that \(\frac{\text{\(h_p^T h_q\)}}{2N_t}\) converges to 0 almost surely. This also implies that \(\frac{\text{\(h_p^T h_q\)}}{2N_t}\) converges in distribution to the constant 0, and hence
by Slutsky’s theorem, $\frac{h_j^T h_k}{||h_j||^2 + ||h_k||^2}$ converges in distribution to 0. Since, if a sequence of r.v’s converges in distribution to a constant then the sequence converges in probability to that constant, we conclude that indeed $\frac{h_j^T h_k}{||h_j||^2 + ||h_k||^2}$ converges in probability to 0. This proves the base case.

Induction Hypothesis: Let $z_{u_n, d} \overset{P}{\to} 0$ as $N_t \to \infty$, $\forall n = 2, 3, \cdots, m$.

Induction Step: Proof for $n = m + 1$: We have

$$z_{u_{(m+1)}, d} = \frac{\sum_{k=1}^{m+1} \sum_{j=k+1}^{m+1} h_j^T h_k d_j d_k}{\sum_{j=1}^{m+1} ||h_j||^2}$$

$$= \frac{\sum_{k=1}^{m} \sum_{j=k+1}^{m} h_j^T h_k d_j d_k + \sum_{k=1}^{m} h_k^T q k d_k}{\sum_{j=1}^{m} ||h_j||^2} + \frac{\sum_{k=1}^{m} h_k^T q k d_k}{\sum_{j=1}^{m} ||h_j||^2}$$

$$= 1 + \frac{||h_{(m+1)}||^2}{\sum_{j=1}^{m} ||h_j||^2}. \quad (B.4)$$

Using Slutsky’s theorem and the strong law of large numbers, it can be shown that the denominator in (B.4) converges to $(1 + \frac{1}{m})$ in probability. Also, from the induction hypothesis, the term $\sum_{k=1}^{m} \sum_{j=k+1}^{m} h_j^T h_k d_j d_k$ in the numerator of (B.4) converges in probability to 0. Therefore, the numerator in (B.4) converges to the same distribution that $\sum_{j=1}^{m} ||h_j||^2$ converges to. Also, the term $\frac{\sum_{k=1}^{m} h_k^T q k d_k}{\sum_{j=1}^{m} ||h_j||^2}$ is the same as $

\frac{(\sum_{k=1}^{m} h_k^T q k d_k) \sum_{j=1}^{m} ||h_j||^2}{(\sum_{j=1}^{m} ||h_j||^2)^2}$. Further, from the strong law of large numbers, the term $(\sum_{j=1}^{m} ||h_j||^2)/(m N_t)$ converges almost surely to 1. Therefore, from Slutsky’s theorem, we know that $\frac{\sum_{k=1}^{m} h_k^T q k d_k}{(\sum_{j=1}^{m} ||h_j||^2)^2} \sum_{j=1}^{m} ||h_j||^2$ converges in distribution to the distribution to which the term $(\sum_{k=1}^{m} h_k^T q k d_k)/(m N_t)$ converges.

For a given vector $d$, $h_{ik} d_{ik}$ is a random vector whose distribution is the same as that of $h_{ik}$. Therefore, applying Lemma 4, we see that the term $(\sum_{k=1}^{m} h_k^T q k d_k)/(m N_t)$ converges almost surely to 0. Hence, the numerator in (B.4) converges in probability to the constant 0. Therefore, $z_{u_{(m+1)}, d} \overset{P}{\to} 0$ as $N_t \to \infty$. This proves the induction step and completes the proof of Lemma 5. $\square$
Appendix C

Conjecture 1, Chapter 3

We present conjecture 1 of Chapter 3 in this appendix. Through the principle of mathematical induction, we conjecture that for any detection algorithm $A_1(.)$ satisfying property (3.13), and any $0 < \delta < 1$, for each $m = 2, 3, \ldots, 2N_t$, there exists an integer $N_m(\delta)$ such that for all $N_t \geq N_m(\delta)$, and any $(x, n), p_m(x, n) > (1 - \delta)$.

For a given $H, x$ and $n$, let $d = A_1(H, y)$. We first present the base case for $m = 2$, where we show that for a given $(x, n)$, if $n \in R_{d^1}$, then $n \in R_{d^2}$ with high probability (w.r.t. the probability distribution of $H$).

**Base Case ($m = 2$):**

Since $A_1(.)$ satisfies property (3.13), it must be true that $d \in \mathbb{D}^1_{H,x,n}$. From (3.12), this further implies that $n \in R_{d^1}$. Therefore, from the definition of $R_{d^m}$, $n$ satisfies (3.6). To prove the base case we need to consider the event $\{n \in R_{d^2}\}$.

For $n$ to belong to $R_{d^2}$, in addition to satisfying (3.6), $n$ must also satisfy the following equation $\forall p, q = 1, \ldots, 2N_t, p \neq q$:

$$
(n + H(x - d) + h_p d_p + h_q d_q)^T (h_p d_p + h_q d_q) \geq 0, \tag{C.1}
$$

which can be rewritten as

$$
(n + H(x - d))^T h_p d_p + (n + H(x - d))^T h_q d_q \geq -\|h_p\|^2 - \|h_q\|^2 - 2d_p d_q h_p^T h_q. \tag{C.2}
$$
Since \( n \) satisfies (3.6), it satisfies the following two equations:

\[
(n + H(x - d))^T h_p d_p \geq -\|h_p\|^2, \\
(n + H(x - d))^T h_q d_q \geq -\|h_q\|^2.
\]  

(C.3)

Comparing (C.3) and (C.2), we notice that if \( h_p \) and \( h_q \) are orthogonal, then \( n \) trivially satisfies (C.2) for all \( N_t \). Therefore, when \( h_p \) and \( h_q \) are non-orthogonal, the only extra term in the RHS of (C.2) is \( 2d_p d_q h_p^T h_q \). Applying Lemma 5 of Chapter 3, with \( n = 2 \), we see that as \( N_t \to \infty \), the r.v. \( \frac{h_p^T h_q}{\|h_p\|^2 + \|h_q\|^2} \) converges to zero in probability. Then, we can write, for any \( \epsilon, 0 \leq \epsilon \leq 1 \)

\[
p \left( \frac{|h_p^T h_q|}{\|h_p\|^2 + \|h_q\|^2} > \epsilon \right) < \epsilon, \quad \forall N_t > f(\epsilon).
\]  

(C.4)

For each pair \( (p, q) \) and a given \( H \) we define the following events

\[
E_1(p, q, H) \triangleq \{ \frac{|h_p^T h_q|}{\|h_p\|^2 + \|h_q\|^2} < \epsilon \} \\
E_2(p, q, H) \triangleq \{ \frac{|h_p^T h_q|}{\|h_p\|^2 + \|h_q\|^2} > \epsilon \} \\
E_3(p, q, H) \triangleq \{ (n + H(x - d) + h_p d_p + h_q d_q)^T (h_p d_p + h_q d_q) \} < 0 \\
E_{11}(p, q, H) \triangleq \{ 0 < h_p^T h_q < \epsilon (\|h_p\|^2 + \|h_q\|^2) \} \\
E_{12}(p, q, H) \triangleq \{ 0 > h_p^T h_q > -\epsilon (\|h_p\|^2 + \|h_q\|^2) \} \\
E_{+1}(p, q, H) \triangleq \{ d_p d_q = +1 \} \\
E_{-1}(p, q, H) \triangleq \{ d_p d_q = -1 \} \\
E_c(p, q, H) \triangleq \left( E_{+1}(p, q, H) \cap E_{11}(p, q, H) \right) \cup \left( E_{-1}(p, q, H) \cap E_{12}(p, q, H) \right).
\]  

(C.5)

For \( n \notin \mathcal{R}_d \), \( E_3(p, q, H) \) must be true for at least some pair \( (p, q) \). Further, let \( E_3(H) \triangleq \)
\[ \cup_{(p,q) \neq q} E_3(p, q, H). \bar{p}_2(x, n) \text{ can now be expressed as} \]

\[ \bar{p}_2(x, n) = \mathbb{E}_H[I(n \notin \mathcal{R}_d \mid H, x, n, d = A_1(H, y))] \]

\[ = \mathbb{E}_H[I(E_3(H) \mid H, x, n, d = A_1(H, y))] \]

\[ \leq \sum_{(p,q) \neq q} \mathbb{E}_H[I(E_3(p, q, H) \mid H, x, n, d = A_1(H, y))]. \quad (C.6) \]

The last inequality follows from the union bound. However, this bound is not very tight. Due to analytical intractability, it is difficult to get a meaningful bound. This is the primary difficulty in proving the conjecture. To gain some insight as to why the conjecture may actually be true, we attempt to bound \( \mathbb{E}_H[I(E_3(p, q, H) \mid H, x, n, d = A_1(H, y))] \) for each \((p, q)\).

\[ \mathbb{E}_H[I(E_3(p, q, H) \mid H, x, n, d = A_1(H, y))] \]

\[ = \mathbb{E}_H[I(E_3(p, q, H), E_1(p, q, H) \mid H, x, n, d = A_1(H, y))] \]

\[ + \mathbb{E}_H[I(E_3(p, q, H), E_2(p, q, H) \mid H, x, n, d = A_1(H, y))] \]

\[ < \mathbb{E}_H[I(E_3(p, q, H), E_1(p, q, H) \mid H, x, n, d = A_1(H, y))] \]

\[ + \mathbb{E}_H[I(E_2(p, q, H) \mid H, x, n, d = A_1(H, y)))] \]

\[ < \mathbb{E}_H[I(E_3(p, q, H), E_1(p, q, H) \mid H, x, n, d = A_1(H, y)))] + \epsilon \quad (C.7) \]

The last inequality follows from \((C.4)\). \( \mathbb{E}_H[I(E_3(p, q, H), E_1(p, q, H) \mid H, x, n, d = A_1(H, y)))] \) can now be expressed as

\[ \mathbb{E}_H[I(E_3(p, q, H), E_1(p, q, H) \mid H, x, n, d = A_1(H, y)))] \]

\[ = \mathbb{E}_H[I(E_3(p, q, H), E_1(p, q, H), E_c(p, q, H) \mid H, x, n, d = A_1(H, y))] \]

\[ + \mathbb{E}_H[I(E_3(p, q, H), E_1(p, q, H), \bar{E}_c(p, q, H) \mid H, x, n, d = A_1(H, y))], \quad (C.8) \]

where \( \bar{E}_c(p, q, H) \) denotes the complementary event of \( E_c(p, q, H) \).
We now make the following important observations. For any pair \((p, q)\), if \(E_c(p, q, H)\) and \(E_1(p, q, H)\) is true, then

\[
- (\|h_p\|^2 + \|h_q\|^2) > - (\|h_p\|^2 + \|h_q\|^2 + 2d_p d_q h^T_p h_q) > - (\|h_p\|^2 + \|h_q\|^2) (1 + 2\epsilon).
\]

(C.9)

Since \(n \in R_d\), \(n\) satisfies (C.3), and hence satisfies the following equation:

\[
(n + H(x - d))^T (h_p d_p + h_q d_q) \geq - (\|h_p\|^2 + \|h_q\|^2).
\]

(C.10)

Using (C.9) and (C.10), we see that \(n\) satisfies (C.2). So, if \(E_c(p, q, H)\) and \(E_1(p, q, H)\) are true, it can be concluded that \(E_3(p, q, H)\) is false. Therefore,

\[
\mathbb{E}_H[I(E_3(p, q, H), E_1(p, q, H), E_c(p, q, H) | H, x, n, d = A_1(H, y))] = 0.
\]

Hence, this implies that

\[
\mathbb{E}_H[I(E_3(p, q, H), E_1(p, q, H) | H, x, n, d = A_1(H, y))]
= \mathbb{E}_H[I(E_3(p, q, H), E_1(p, q, H), \overline{E}_c(p, q, H) | H, x, n, d = A_1(H, y))].
\]

(C.11)

We next analyze \(\mathbb{E}_H[I(E_3(p, q, H), E_1(p, q, H), \overline{E}_c(p, q, H) | H, x, n, d = A_1(H, y))]\). It can be shown that if the composite event \(\{E_3(p, q, H), E_1(p, q, H), \overline{E}_c(p, q, H)\}\) is true, it implies that the following equations are satisfied.

\[
- (\|h_p\|^2 + \|h_q\|^2) (1 - 2\epsilon) > - (\|h_p\|^2 + \|h_q\|^2 + 2d_p d_q h^T_p h_q) > - (\|h_p\|^2 + \|h_q\|^2)
\]

\[
(n + H(x - d))^T h_p d_p + (n + H(x - d))^T h_q d_q \leq -\|h_p\|^2 - \|h_q\|^2 - 2d_p d_q h^T_p h_q.
\]

(C.12)

Combining the two sub equations in (C.12), we have

\[
(n + H(x - d))^T h_p d_p + (n + H(x - d))^T h_q d_q \leq - (\|h_p\|^2 + \|h_q\|^2) (1 - 2\epsilon).
\]

(C.13)
Also, since \(d \in D_{x,n,H}^1\), we can rewrite (C.13) as

\[-(\|h_p\|^2 + \|h_q\|^2) \leq (n + H(x - d))^T(h_p d_p + h_q d_q) \leq -(\|h_p\|^2 + \|h_q\|^2)(1 - 2\epsilon)\].

(C.14)

So, it can be stated that if the composite event \(\{E_3(p,q,H), E_1(p,q,H), \bar{E}_c(p,q,H)\}\) is true, it implies (C.14). Let us define the event \(E_4(p,q,H) \triangleq \{-(\|h_p\|^2 + \|h_q\|^2) \leq (n + H(x - d))^T(h_p d_p + h_q d_q) \leq -(\|h_p\|^2 + \|h_q\|^2)(1 - 2\epsilon)\}\).

Therefore, \(I(E_3(p,q,H), E_1(p,q,H), \bar{E}_c(p,q,H) \mid H, x, n, d = A_1(H,y)) = 1\) implies that \(I(E_4(p,q,H) \mid H, x, n, d = A_1(H,y))) = 1\). Hence,

\[
\mathbb{E}_H[I(E_3(p,q,H), E_1(p,q,H), \bar{E}_c(p,q,H) \mid H, x, n, d = A_1(H,y)))] < \mathbb{E}_H[I(E_4(p,q,H) \mid H, x, n, d = A_1(H,y))].
\]

We next define a function \(f_2(\epsilon), \epsilon > 0,\) as

\[
f_2(\epsilon) \triangleq \mathbb{E}_H[I(E_4(p,q,H) \mid H, x, n, d = A_1(H,y))].
\]

(C.16)

Further,

\[
\lim_{\epsilon \to 0} f_2(\epsilon) = 0,
\]

(C.17)

since for each \((p,q), \) as \(\epsilon \to 0,\) the upper bound on \((n + H(x - d))^T(h_p d_p + h_q d_q)\) in (C.14) approaches the lower bound. Similarly, it can also be shown that, \(f_2(\epsilon)\) is a monotonically increasing function of \(\epsilon.\) Using (C.16), (C.15) can be written as

\[
\mathbb{E}_H[I(E_3(p,q,H), E_1(p,q,H), \bar{E}_c(p,q,H) \mid H, x, n, d = A_1(H,y)))] < f_2(\epsilon).
\]

(C.18)
Using (C.11) and (C.18), (C.7) can be written as

\[ \mathbb{E}_H [I(E_3(p, q, H) | H, x, n, d = A_1(H, y))] < f_2(\epsilon) + \epsilon. \]  

(C.19)

Now, define \( g_2(\epsilon) \triangleq f_2(\epsilon) + \epsilon \). \( g_2 \) is a monotonic function and therefore has an inverse \( g_2^{-1} \). Hence, for any \( \delta > 0 \), there exists an \( \epsilon \) such that \( \epsilon = g_2^{-1}(\delta) \). Using (C.4) and the above definitions, we can write that

\[
N_t > f(\epsilon) \\
> f(g_2^{-1}(\delta)) \\
> N_2(\delta),
\]

(C.20)

where \( N_2 \triangleq f \circ g_2^{-1} \). Using (C.19), and the above definitions we can state that

\[ \mathbb{E}_H [I(\bar{E}_3(p, q, H) | H, x, n, d = A_1(H, y))] > 1 - \delta. \]  

(C.21)

Hence, we have shown that for any \( \delta, 0 \leq \delta \leq 1 \), there exists an integer \( N_2(\delta) \) such that for \( N_t > N_2(\delta) \), any arbitrary pair \((p, q)\) satisfies (C.2) with a high probability of \((1 - \delta)\).

For \( n \in \mathcal{R}_{d^2} \), (C.2) should be satisfied for all possible pairs \((p, q)\). Since we have shown that indeed (C.2) is satisfied by any arbitrary pair with high probability, we conjecture that (C.2) would be satisfied by all pairs with high probability.

**Induction Hypothesis:** For any detection algorithm satisfying property (3.13), and any \( 0 < \delta < 1 \), there exists an integer \( N_{m-1}(\delta) \) such that for all \( N_t \geq N_{m-1}(\delta) \), and any \((x, n)\), \( p_{m-1}(x, n) > (1 - \delta) \). In the induction hypothesis, we basically assume that if \( n \in \mathcal{R}_{d^1} \), then \( n \in \mathcal{R}_{d^{m-1}} \) with high probability.

**Induction Step:** We conjecture that for any \( 0 < \delta < 1 \), there exists an integer \( N_m(\delta) \) such that for all \( N_t \geq N_m(\delta) \), and any \((x, n)\), \( p_m(x, n) > (1 - \delta) \). Based on the definition of \( p_m(x, n) \), we conjecture that \( n \in \mathcal{R}_{d^m} \) with high probability. For \( n \) to belong to \( \mathcal{R}_{d^m} \), \( n \)
must satisfy the following equation for all possible \(m\)-tuples \((i_1, i_2, \ldots, i_m)\):

\[
(n + H(x - d) + \left(\sum_{j=1}^{m} h_{ij} d_{ij}\right))^T \left(\sum_{j=1}^{m} h_{ij} d_{ij}\right) \geq 0,
\]

(C.22)

which can be written as

\[
(n + H(x - d))^T \left(\sum_{j=1}^{m-1} h_{ij} d_{ij}\right) + (n + H(x - d))^T h_{im} d_{im}
\geq -\|\sum_{j=1}^{m-1} h_{ij} d_{ij}\|^2 - \|h_{im}\|^2 - 2 \left(\sum_{j=1}^{m-1} h_{ij} d_{ij}\right)^T h_{im} d_{im}.
\]

(C.23)

However, we know from the induction hypothesis that \((n + H(x - d))^T (\sum_{j=1}^{m-1} h_{ij} d_{ij}) \geq -\|\sum_{j=1}^{m-1} h_{ij} d_{ij}\|^2\) with high probability. Also, since \(n \in \mathcal{R}_d\), we know that \((n + H(x - d))^T h_{im} d_{im} \geq -\|h_{im}\|^2\). Therefore, if the term \(2 (\sum_{j=1}^{m-1} h_{ij} d_{ij})^T h_{im} d_{im}\) in the RHS of (C.23) were 0, then (C.22) would have been trivially satisfied with high probability. We now show that the contribution of the term \(2 (\sum_{j=1}^{m-1} h_{ij} d_{ij})^T h_{im} d_{im}\) when compared to the other two terms in the RHS (C.23) converges to 0 as \(N_t \to \infty\).

Define a r.v. \(v_m \triangleq \frac{2 (\sum_{j=1}^{m-1} h_{ij} d_{ij})^T h_{im} d_{im}}{\|\sum_{j=1}^{m-1} h_{ij} d_{ij}\|^2 + \|h_{im}\|^2}\). Our objective is to show that as \(N_t \to \infty\), \(v_m \to 0\) in probability. This is equivalent to proving that \(w_m \triangleq v_m + 1 = \frac{\|\sum_{j=1}^{m-1} h_{ij} d_{ij}\|^2}{\|h_{im}\|^2 + \|\sum_{j=1}^{m-1} h_{ij} d_{ij}\|^2}\) converges to one in probability as \(N_t \to \infty\). We can write \(w_m\) as

\[
w_m = \frac{\|\sum_{j=1}^{m-1} h_{ij} d_{ij}\|^2}{\sum_{j=1}^{m} \|h_{ij}\|^2 \left(\frac{\|h_{im}\|^2}{\|h_{im}\|^2 + \|\sum_{j=1}^{m} |h_{ij}|^2\|^2}\right)}.
\]

(C.24)

From Lemma 5 of Chapter 3, we know that for any integer \(m, 1 \leq m \leq 2N_t\), it is true that \(\frac{\sum_{k=1}^{m} \sum_{j=k+1}^{m} h_{ik}^T h_{ij} d_{ik} d_{ij}}{\sum_{j=1}^{m} \|h_{ij}\|^2}\) converges to 0 in probability as \(N_t \to \infty\). By Slutsky’s theorem, this is equivalent to

\[
2 \sum_{k=1}^{m} \sum_{j=k+1}^{m} h_{ik}^T h_{ij} d_{ik} d_{ij} \sum_{j=1}^{m} \|h_{ij}\|^2 + 1 = \frac{\|\sum_{j=1}^{m} h_{ij} d_{ij}\|^2}{\sum_{j=1}^{m} \|h_{ij}\|^2} \to 1
\]

as \(N_t \to \infty\). We shall use this result to prove the convergence of \(w_m\) in (C.24). Using
(C.25), it can be seen that the numerator of \( \omega_m \) in (C.24) converges to 1 as \( N_t \to \infty \), i.e.,

\[
\frac{\left\| \sum_{j=1}^{m} h_{ij} d_{ij} \right\|^2}{\sum_{j=1}^{m} \left\| h_{ij} \right\|^2} \overset{p}{\longrightarrow} 1, \quad \text{as } N_t \to \infty. \tag{C.26}
\]

In the denominator of (C.24), it can be shown that the term

\[
\frac{\left\| h_{im} \right\|^2}{\sum_{j=1}^{m} \left\| h_{ij} \right\|^2} \overset{p}{\longrightarrow} \frac{1}{m}, \quad \text{as } N_t \to \infty. \tag{C.27}
\]

The 2nd term in the denominator of (C.24) can be rewritten as

\[
\frac{\left\| \left( \sum_{j=1}^{m-1} h_{ij} d_{ij} \right) \right\|^2}{\sum_{j=1}^{m} \left\| h_{ij} \right\|^2} = \frac{\frac{\left\| \sum_{j=1}^{m-1} h_{ij} d_{ij} \right\|^2}{\sum_{j=1}^{m} \left\| h_{ij} \right\|^2}}{\sum_{j=1}^{m} \left\| h_{ij} \right\|^2 + 1}. \tag{C.28}
\]

Similar to the derivation of (C.25), we can claim that the numerator in (C.28) converges to one in probability. From Slutsky’s theorem, it can be shown that \( \frac{\left\| h_{im} \right\|^2}{\sum_{j=1}^{m} \left\| h_{ij} \right\|^2} \) converges to \( \frac{1}{m-1} \) in probability. Using this and Slutsky’s theorem, it can be shown that (C.28) converges to \( \frac{m-1}{m} \) in probability. Using this result along with (C.26),(C.27) and Slutsky’s theorem in (C.24), it can be shown that \( \omega_m \) converges to one in probability as \( N_t \to \infty \). This, therefore, implies that \( \nu_m \) converges to zero in probability. Following along the same lines as the conjecture for the base case, it can be conjectured that for any \( \delta, 0 \leq \delta \leq 1 \), there exists an integer \( N_m(\delta) \) such that for \( N_t > N_m(\delta), p_m(x, n) > 1 - \delta \) (i.e. \( n \in \mathcal{R}_d^m \) with high probability). This completes the induction step. \( \square \)
Appendix D

Proof of Theorem 1, Chapter 5

Towards proving Theorem 1, we shall find the following Lemma useful (see Proposition 1 in [101] for the same lemma and its proof).

**Lemma 1.** Given a real scalar channel modeled by $y = \sqrt{\alpha}x + n$, where $x = \pm \sqrt{E_s}$, $n \sim \mathcal{N}(0, \sigma^2)$. Let $F(\alpha) = C\alpha^k + o(\alpha^k)$, for $\alpha \to 0^+$ be the cdf (cumulative density function) of $\alpha$, where $C$ is a constant and $k$ is a positive integer\footnote{Any function $f(x)$ in a single variable $x$ is said to be $o(g(x))$, i.e., $f(x) = o(g(x))$, if $\lim_{x \to 0} \frac{f(x)}{g(x)} \to 0$.}. Let $\gamma = E_s/\sigma^2$ be the SNR. Then the probability of error is given by $P(\gamma) = \mathbb{E}_\alpha[Q(\sqrt{\alpha\gamma})]$. The asymptotic error probability for $\gamma \to \infty$ is given by

$$P = \frac{C((2k - 1) \cdot (2k - 3) \cdot 5 \cdot 3 \cdot 1)}{2} \gamma^{-k} + o(\gamma^{-k}). \quad (D.1)$$

The proof of Theorem 1 is as follows. Let $\{\Re(u_k) \to \Re(v_k)\}$ denote the pairwise error event that, given $u_k$ was transmitted on the $k$th pair, the real part of the ML detector for the $k$th pair decodes in favor of some other vector $\Re(v_k)$. Further, let us denote the probability of this event (i.e., PEP) by $P'_k(\Re(u_k) \to \Re(v_k))$. Using the union bounding technique, $P'_k(\Re(u_k))$ is then upper bounded by the sum of all the possible pairwise error probabilities. From (5.22), it is clear that this upper bound on $P'_k(\Re(u_k))$ induces.
an upper bound on $P'_k$, which is given by

$$P'_k \leq \frac{1}{|S_k|} \sum_{\mathbb{R}(u_k) \neq \mathbb{R}(v_k)} \sum_{\mathbb{R}(v_k)} P'_k(\mathbb{R}(u_k) \rightarrow \mathbb{R}(v_k)).$$

(D.2)

Due to Gaussian noise, this can be further written as

$$P'_k \leq \frac{1}{|S_k|} \sum_{\mathbb{R}(u_k) \neq \mathbb{R}(v_k)} \sum_{\mathbb{R}(v_k)} \mathbb{E}\left[ Q\left( \sqrt{\frac{d^2_k(\mathbb{R}(u_k), \mathbb{R}(v_k), A_k)}{2N_0}}\right) \right].$$

(D.3)

where $d^2_k(\mathbb{R}(u_k), \mathbb{R}(v_k), A_k) \triangleq \|M_k(\mathbb{R}(u_k) - \mathbb{R}(v_k))\|^2$, and $Q(x) \triangleq \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$. The expectation in (D.3) is over the joint distribution of the channel gain $(\lambda_{i_k}, \lambda_{j_k})$. The joint pdf (probability density function) of the ordered eigenvalues of $H^H H$, i.e., $\{\lambda^2_1 > \lambda^2_2 \cdots > \lambda^2_{N_r}\}$ is given by the well known Wishart distribution [104]. However, in (D.3) evaluating the expectation over $(\lambda_{i_k}, \lambda_{j_k})$ is still a difficult problem except for trivial cases (like $N_t = N_r = 2$). We, therefore, try to bound $d^2_k(\mathbb{R}(u_k), \mathbb{R}(v_k), A_k)$ such that the bound depends only on $\lambda_{i_k}$. Since $\lambda_{i_k} \geq \lambda_{j_k} \geq 0$, using the definition of $M$ and $M_k$ we have

$$d^2_k(\mathbb{R}(u_k), \mathbb{R}(v_k), A_k) \geq \lambda^2_{i_k} \tilde{d}^2_k(\mathbb{R}(u_k), \mathbb{R}(v_k), A_k),$$

(D.4)

where $\tilde{d}^2_k(\mathbb{R}(u_k), \mathbb{R}(v_k), A_k) \triangleq e^2_{k,1}, e_k \triangleq A_k(\mathbb{R}(u_k) - \mathbb{R}(v_k))$. $e_{k,1}$ denotes the first component of the 2-dimensional vector $e_k$. We further define the generalized minimum distance as follows:

$$g_k(A_k) = \min_{\mathbb{R}(u_k) \neq \mathbb{R}(v_k)} \tilde{d}^2_k(\mathbb{R}(u_k), \mathbb{R}(v_k), A_k).$$

(D.5)

Since $Q(.)$ is a monotonically decreasing function with increasing argument, we can
further upper bound (D.3) using (D.5) as follows:

\[
P'_k \leq (|S_k| - 1) \mathbb{E} \left[ Q \left( \sqrt{\frac{\lambda^2 g_k(A_k)}{2N_0}} \right) \right]. \tag{D.6}
\]

For a Rayleigh faded channel, the marginal pdf of the \(s\)th eigenvalue \(\lambda_s^2\) (for \(\lambda_s^2 \to 0\)) is given by [105]

\[
f(\lambda_s^2) = C(s)(\lambda_s^2)^{N_t(s)N_r(s)-1} + o\left((\lambda_s^2)^{N_t(s)N_r(s)-1}\right), \tag{D.7}
\]

where \(N_t(s) \triangleq (N_t - s + 1), N_r(s) \triangleq (N_r - s + 1)\) and \(C(s)\) is a constant given in [105]. Using the pdf in (D.7), the cdf \(F_s(u) = P(\lambda_s^2 \leq u)\) (for \(u \to 0^+\)) is given by

\[
F_s(u) = D(s)u^{N_t(s)N_r(s)} + o\left(u^{N_t(s)N_r(s)}\right), \tag{D.8}
\]

where \(D(s) \triangleq \frac{C(s)}{N_t(s)N_r(s)}\). Using Lemma 1 and (D.8), the bound in (D.6) can be further written as

\[
P'_k \leq (|S_k| - 1)c_k \left(\frac{\gamma g_k(A_k)}{2P_T}\right)^{-\delta_k} + o(\gamma^{-\delta_k}) \tag{D.9}
\]

where \(\delta_k \triangleq (N_t - i_k + 1)(N_r - i_k + 1)\) and \(c_k \triangleq \frac{C(i_k)((2\delta_k - 1) \cdots 5.3.1)}{2\delta_k}.\) \hfill \blacksquare
Appendix E

Proof of Theorem 2, Chapter 5

Let \( d(\theta_k, \lambda_{i_k}, \lambda_{j_k}) \triangleq \min_{(p,q) \in \mathbb{S}_2} d_k^2(p, q, \theta_k) \), where \( d_k^2(p, q, \theta_k) \) is defined in (5.50). The objective is to find the optimal \( \theta_k \) which maximizes \( d(\theta_k, \lambda_{i_k}, \lambda_{j_k}) \). The set \( \mathbb{S}_2 \) contains exactly 8 elements. Also, due to sign symmetry of this set (i.e., if \( (p, q) \in \mathbb{S}_2 \), then so do \( (p, -q) \), \( (-p, q) \) and \( (-p, -q) \)), there are actually only 4 distances to be computed. For a given angle \( \theta_k \), these distances are enumerated as follows:

\[
\begin{align*}
d_1 &= \lambda_{i_k}^2 \cos^2(\theta_k) + \lambda_{j_k}^2 \sin^2(\theta_k) \\
d_2 &= \lambda_{i_k}^2 \sin^2(\theta_k) + \lambda_{j_k}^2 \cos^2(\theta_k) \\
d_3 &= \lambda_{i_k}^2 (\cos(\theta_k) + \sin(\theta_k))^2 + \lambda_{j_k}^2 (\cos(\theta_k) - \sin(\theta_k))^2 \\
d_4 &= \lambda_{i_k}^2 (\cos(\theta_k) - \sin(\theta_k))^2 + \lambda_{j_k}^2 (\cos(\theta_k) + \sin(\theta_k))^2.
\end{align*}
\]

Therefore, \( d(\theta_k, \lambda_{i_k}, \lambda_{j_k}) \) can be expressed in terms of these distances as

\[
d(\theta_k, \lambda_{i_k}, \lambda_{j_k}) = \min(d_1, d_2, d_3, d_4)
= \min(\min(d_1, d_2), \min(d_3, d_4)).
\]

It can be shown that for the maximization in (5.49), it suffices to only consider the range of \( \theta_k \) to be \([0, \pi/4]\). Also, it is given that \( \lambda_{i_k} \geq \lambda_{j_k} \) due to the order of the singular values in the SVD decomposition and the way in which pairing is done for X-Codes. Using
these two facts, it can be concluded that

\[ d_2 = \min(d_1, d_2) \quad \text{(E.6)} \]
\[ d_4 = \min(d_3, d_4), \quad \text{(E.7)} \]

and therefore

\[ d(\theta_k, \lambda_{i_k}, \lambda_{j_k}) = \min(d_2, d_4). \quad \text{(E.8)} \]

Let

\[ \beta_k \triangleq \frac{\lambda_{i_k}}{\lambda_{j_k}}. \quad \text{(E.9)} \]

Then, \( d_4 \) is the minimum iff the following condition is satisfied:

\[ \frac{1}{\beta_k^2} \leq h(\theta_k), \quad \text{(E.10)} \]

where

\[ h(\theta_k) \triangleq 1 - \frac{1}{\sin(2\theta_k) + \sin^2(\theta_k)}. \quad \text{(E.11)} \]

It can be seen that over the interval \((0, \frac{\pi}{4}]\), \( h(\theta_k) \) is a continuous and monotonically increasing function. This is because the first derivative w.r.t \( \theta_k \) exists and is always positive. The maximum value of \( h(\theta_k) \) over this interval is \( \frac{1}{3} \). Therefore, we now consider two situations depending upon whether \( \beta_k \) is greater than or less than \( \sqrt{3} \).

If \( \beta_k \leq \sqrt{3} \), then \( \frac{1}{\beta_k^2} \geq \frac{1}{3} \). Since \( h(\theta_k) \) is always less than \( \frac{1}{3} \), we can conclude that the condition in (E.10) is never satisfied, and therefore \( d_2 \) is the minima. Further, since \( \lambda_{i_k} \geq \lambda_{j_k}, d_2 \) is a monotonically increasing function of \( \theta_k \), and therefore the solution to the maximization problem in (5.49) is \( \frac{\pi}{4} \).

If \( \beta_k \geq \sqrt{3} \), then \( \frac{1}{\beta_k^2} \leq \frac{1}{3} \). Since \( h(\theta_k) \) is a monotonically increasing function, we observe
that \( d_4 \) is the minima when \( \theta_k \geq \theta_k^* \); else \( d_2 \) is the minima. Here, \( \theta_k^* \) is such that \( \theta_k^* \in [0 \frac{\pi}{4}] \) and \( h(\theta_k^*) = \frac{1}{\beta_k} \). Further, it is observed that \( d_2 \) is a monotonically increasing function of \( \theta_k \), whereas \( d_4 \) is monotonically decreasing. Also, \( d_2 = d_4 \) when \( \theta_k = \theta_k^* \). Therefore, it can be concluded that \( \min(d_2, d_4) \) is maximized when \( \theta_k = \theta_k^* \). Hence, for \( \beta_k \geq \sqrt{3} \), the solution to the maximization problem in (5.49) is \( \theta_k^* \). We now solve for \( \theta_k^* \). Using the definition of \( h(\theta_k) \), we have
\[
\frac{1}{\beta_k^2} = 1 - \frac{1}{\sin(2\theta_k^*) + \sin^2(\theta_k^*)},
\]
(E.12)
or equivalently
\[
\tan^2(\theta_k^*) - 2(\beta_k^2 - 1) \tan(\theta_k^*) + \beta_k^2 = 0.
\]
(E.13)
The last equation is quadratic in \( \tan(\theta_k^*) \), and the solution which results in \( \theta_k^* \in [0 \frac{\pi}{4}] \) is given by
\[
\theta_k^* = \tan^{-1}\left[ (\beta_k^2 - 1) - \sqrt{((\beta_k^2 - 1)^2 - \beta_k^2)} \right].
\]
(E.14)
Combining the optimal angles obtained for \( \beta_k \leq \sqrt{3} \) and \( \beta_k > \sqrt{3} \), we get the solution to the maximization problem as stated in (5.51).
Appendix F

Proof of Theorem 3, Chapter 5

We first get an expression for $d_{k,\min}^2(a_k, b_k)$ as defined in (5.74). For any code vector at index $v$ which is even, the nearest distance to any other code vector with even index is $4\lambda^2 i_k a_k^2$. The nearest distance to any code vector at odd index is $\lambda^2 i_k a_k^2 + 4\lambda^2 j_k b_k^2$. The same holds true if $v$ is odd. Hence, $d_{k,\min}^2(a_k, b_k)$ is given by

$$d_{k,\min}^2(a_k, b_k) = \min\left(4\lambda^2 i_k a_k^2, \lambda^2 i_k a_k^2 + 4\lambda^2 j_k b_k^2\right). \quad (F.1)$$

Therefore, our objective is to solve the following constrained min-max optimization problem:

$$(a_k^*, b_k^*) = \arg \max_{(a_k, b_k) \in \mathbb{R}^2 \mid b_k^2 + a_k^2 \frac{M^2-1}{12} = \frac{P_T}{N_r}} \min\left(4\lambda^2 i_k a_k^2, \lambda^2 i_k a_k^2 + 4\lambda^2 j_k b_k^2\right). \quad (F.2)$$

The first argument of the $\min(.,.)$ function, $T_1 \triangleq 4\lambda^2 i_k a_k^2$ is geometrically a straight line w.r.t $a_k^2$ passing through the origin and attaining a maximum value of $\frac{48P_T\lambda^2 i_k}{N_r(M^2-1)}$ at $a_k^2 = \frac{12P_T}{N_r(M^2-1)}$ (due to the transmit power constraint $a_k^2 \leq \frac{12P_T}{N_r(M^2-1)}$). The second argument of the $\min(.,.)$ function is $T_2 \triangleq (\lambda^2 i_k a_k^2 + 4\lambda^2 j_k b_k^2)$. Since $b_k^2 + a_k^2 \frac{(M^2-1)}{12} = \frac{P_T}{N_r}$, we can express $T_2$ as

$$T_2 = \lambda^2 j_k \left(4\frac{P_T}{N_r} + a_k^2 \left(\beta_k^2 - \frac{M^2-1}{3}\right)\right). \quad (F.3)$$
From (F.3) we observe that, if \( \beta_k^2 \geq \frac{M^2-1}{3} \), then \( T_2 \) is a straight line with positive slope, with a value of \( 4\lambda_i^2 \frac{P_T}{N_r} \) at \( a_k = 0 \) and attaining a maximum value of \( \frac{12P_T\lambda_i^2}{N_r(M^2-1)} \) at \( a_k^2 = \frac{12P_T}{N_r(M^2-1)} \). This maximum value is less than the maximum attained by \( T_1 \). Since both \( T_1 \) and \( T_2 \) have positive slopes, the minimum among \( T_1 \) and \( T_2 \) is maximized at \( a_k^2 = \frac{12P_T}{N_r(M^2-1)} \), which implies that \( b_k^* = 0 \). The value of \( d_{k,\min}^2(a_k, b_k) \) at \( a_k = a_k^* \) is the maximum value attained by \( T_2 \). Therefore, when the channel condition exceeds a certain threshold, it is optimal to allocate all power to the stronger channel only.

On the other hand, if \( \beta_k^2 < \frac{M^2-1}{3} \), then \( T_2 \) is a straight line with negative slope, whereas \( T_1 \) has positive slope, and therefore the minimum between them is maximized when they are both equal. Therefore, the optimal \( (a_k^*, b_k^*) \) must satisfy

\[
4\lambda_i^2 a_k^* = \lambda_i^2 a_k^* + 4\lambda_j^2 b_k^*.
\]

Using the fact that \( b_k^* + a_k^* \frac{M^2-1}{12} = \frac{P_T}{N_r} \), the optimal \( (a_k^*, b_k^*) \) is given by

\[
(a_k^*, b_k^*) = \left( \sqrt{\frac{4P_T}{3N_r}}, \sqrt{\frac{P_T}{N_r}} \frac{\beta_k}{\beta_k^2 + \frac{M^2-1}{9}} \right).
\]

Using (F.5), the optimal value of \( d_{k,\min}^2(a_k, b_k) \) for \( \beta_k^2 < \frac{M^2-1}{3} \) is given by

\[
d_{k,\min}^2(a_k^*, b_k^*) = \frac{16P_T\lambda_i^2}{N_r \left( 3\beta_k^2 + \frac{(M^2-1)}{3} \right)}.
\]

\[\blacksquare\]
List of Publications from this Thesis

Journal Papers


near-MAP decoding of large non-orthogonal STBCs using PDA,” journal paper under preparation.

### Conference Papers


Bibliography


effect on the capacity of multi-antenna systems,” IEEE Trans. Commun., vol. 48,

channels with nonzero mean in spacetime coded systems,” IEEE Trans. Signal Pro-

for MIMO systems,” IEEE Trans. Wireless Commun., vol. 5, no. 12, pp. 3579-3587,
December 2006.

coded systems with correlated Rayleigh fading channels using convex optimiza-

in Space-Time Wireless Systems: From Array Processing to MIMO Communica-
tions, edited by H. Bölcskei, D. Gesbert, C. B. Papadias, and Alle-Jan van der Veen,

estimation and data detection in MIMO communication systems,” IEEE Commun.

tive channel estimation and data detection in frequency-selective fading MIMO
2004.


http://www.tsc.uc3m.es/~fernando/power_allocation20090915.pdf


