KTH Electrical Engineering

Bit loading and precoding for MIMO communication systems

SVANTE BERGMAN

Doctoral Thesis in Telecommunications
Stockholm, Sweden 2009

KTH, School of Electrical Engineering
Signal Processing Laboratory
SE-100 44 Stockholm
SWEDEN

TRITA-EE 2009:031
ISSN 1653-5146
ISBN 978-91-7415-359-0

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen i telekommunikation fredagen den 12 juni 2009 klockan 13:15 i hörsal L1, Drottning Kristinas väg 30, Stockholm.

Tryck: Universitetsservice US AB

Abstract

This thesis considers the joint design of bit loading, precoding and receive filters for a multiple-input multiple-output (MIMO) digital communication system. Both the transmitter and the receiver are assumed to know the channel matrix perfectly. It is well known that, for linear MIMO transceivers, orthogonal transmission (i.e., diagonalization of the channel matrix) is optimal for some criteria such as maximum mutual information. It has been shown that if the receiver uses the linear minimum mean squared error (MMSE) detector, the optimal transmission strategy is to perform bit loading on orthogonal subchannels.

In the first part of the thesis, we consider the problem of designing the transceiver in order to minimize the probability of error given maximum likelihood (ML) detection. A joint bit loading and linear precoder design is proposed that outperforms the optimal orthogonal transmission. The design uses lattice invariant operations to transform the channel matrix into a lattice generator matrix with large minimum distance separation at a low price in terms of transmit power. With appropriate approximations, it is shown that this corresponds to selecting lattices with good sphere-packing properties. An algorithm for this power minimization is presented along with a lower bound on the optimization. Apparently, given the optimal ML detector, orthogonal subchannels are (in general) suboptimal.

The ML detector may suffer from high computational complexity, which motivates the use of the suboptimal but less complex MMSE detector. An intermediate detector in terms of complexity and performance is the decision feedback (DF) detector. In the second part of the thesis, we consider the problem of joint bit loading and precoding assuming the DF detector. The main result shows that for a DF MIMO transceiver where the bit loading is jointly optimized with the transceiver filters, orthogonal transmission is optimal. As a consequence, inter-symbol interference is eliminated and the DF part of the receiver is actually not required, only the linear part is needed. The proof is based on a relaxation of the discrete set of available bit rates on the individual subchannels to the set of positive real numbers. In practice, the signal constellations are discrete and the optimal relaxed bit loading has to be rounded. It is shown that the loss due to rounding is small, and an upper bound on the maximum loss is derived. Numerical results are presented that confirm the theoretical results and demonstrate that orthogonal transmission and the truly optimal DF design perform almost equally well. An algorithm that

makes the filter design problem especially easy to solve is presented.

As a byproduct from the work on decision feedback detectors we also present some work on the problem of optimizing a Schur-convex objective under a linearly shifted, or skewed, majorization constraint. Similar to the case with a regular majorization constraint, the solution is found to be the same for the entire class of cost functions. Furthermore, it is shown that the problem is equivalent to identifying the convex hull under a simple polygon defined by the constraint parameters. This leads to an algorithm that produces the exact optimum with linear computational complexity. As applications, two unitary precoder designs for MIMO communication systems that use heterogenous signal constellations and employ DF detection at the receiver are presented.

Acknowledgments

Foremost, I would like to thank my advisor Prof. Björn Ottersten for giving me the opportunity to pursuit a Ph.D. within his field. His support, guidance and encouragement has inspired me in my research, and helped me to face the challenges on the way. He has provided me with an excellent (world class) research environment through the Signal Processing group and the GST.

During my studies, I have also had the opportunity to receive advice from several other distinguished researchers and mentors. It has been a privilege to work with Prof. Daniel Palomar at the Hong Kong University of Technology. He allowed me to visit his group for four months, his guidance and knowhow regarding transceiver design has been invaluable. I would also like to thank Prof. Mats Bengtsson, Prof. Eduard Jorswieck and Dr. Joakim Jaldén for taking the time to discuss research related problems with me.

Research is definitely more fun when working in teams. I would like to thank Simon Järmyr, Dr. Cristoff Martin, Niklas Jaldén, for their cooperation on various papers and projects. I am indebted to my colleagues who have helped to proofread parts of this thesis: Peter von Wrycza, Emil Björnson, Simon Järmyr, John-Olof Nilsson, Petter Wirfält, Dave Zachariah and Dr. Bhavani Shankar. Philosophical discussions with colleagues is an invaluable source of inspiration. I thank the members and alumni of the Signal Processing and Communication Theory groups for the fun and creative atmosphere at floor 4. The always-excellent administrative support has been highly appreciated, thank you Karin Demin and Annika Augustsson.

I am grateful to Prof. Timothy Davidson for taking the time to act as faculty opponent, Prof. Wolfgang Utschick, Prof. Erik Larsson and Prof. Erik Aurell for acting as grading committee.

Although research is mostly great fun, five years of MIMO is not entirely risk free. Without the love and support from my friends and family I would probably have gone nuts by now, so thank you all for maintaining my sanity! Finally, I would like to thank my precious fiancée Frida for believing in me, for her patience, and for reminding me during tough periods that, after all, love is all we need.

Svante Bergman
Stockholm, May 2009

Contents

Chapter 1

Introduction

The recent developments in information and communication technologies have been quite astonishing, even in a historical perspective. Most people today carry mobile phones that allow us unlimited access to anyone anytime at a very low cost. From all inhabited parts of the world, with just a click on a button, we can access most of the worlds written texts, news, recorded songs, films, all in a matter of seconds or minutes. In some sense, wireless access to the internet has redefined the notion of knowledge; to know is no longer to learn and remember, to know is to understand what to look for.

What we have been (and are still) experiencing in terms of new practical applications of technological innovations, is perhaps only comparable to what people during the industrial revolution of the 19'th century may have experienced. In fact, it was in the 19'th century that the first steps towards the modern digital communication systems were taken with the invention of the electrical telegraph. The rate of communication in an electrical telegraph was very much limited by the persons that operated the system. The telegraphist needed a good sense of rhythm and an alert mind in order to transmit or receive messages at a high rate. With the introduction of electronics and computers, the human factor on the rate of communication was no longer the main limitation. The new bottleneck for communication of data was instead given by the physical electromagnetical characteristics of the channel, in particular the signal to noise ratio.

It is partly the processing power provided by the computers that has driven, as well as enabled, the recent dramatically increased usage of digital communications. In the last two decades the prices on advanced communication devices has reduced sharply, while in parallel the operators have increased the coverage and efficiency of the communication networks. Since new sophisticated wireless user terminals are able to present increasingly advanced content, each user is more frequently active and has an incentive to consume more data traffic. As wireless access becomes a necessity for more people, the demand for even better coverage and reliability will grow.

The challenge of supporting more users, increased traffic, better coverage and higher reliability is not a problem the operators can solve solely by installing more infrastructure — the radio spectrum is a finite resource that is strictly regulated and is typically very costly to acquire. For this reason, much research effort has been (and is being) put on advanced signal processing techniques for making use of the available spectrum as efficiently as possible. The aim of this thesis is to provide a contribution to this important scientific field.

1.1 Making the most out of the spectrum

Information theory [CT91], the mathematical theory on how to optimally store and send information, took a great leap in 1948 with the pioneering work by Shannon [Sha48]. Shannon showed that a communication channel, such as a radio link or a magnetic tape, can convey information without errors up to a limit; the channel capacity. Equally important, he showed that it is impossible to send error-free information at a rate above the channel capacity. The capacity is determined by the signal to noise ratio on the channel and it sets a fundamental limit on the rate of communication.

In order to attain data rates close to the channel capacity, the transmitter needs to accumulate the information and encode it using very long codewords. The receiver of the information can then decode the data once the entire codeword has been received. In information theoretic work these codewords are typically infinitely long when transmitting at the rate of the capacity. In practice finite codewords are used, but with a penalty that there is a small but non-zero probability of decoding errors [Gal62, RU08]. Roughly speaking, given that our codebooks are wisely designed, the larger chunks of information that are encoded (using longer codewords), the lower probability of a decoding error is attained. The protection of data against errors by means of coding is commonly referred to as channel coding. One consequence of channel coding is that some delay is inevitable in order to maximize the throughput on the available spectrum.

1.2 The multiple-input multiple-output system

Any communication system transmitting and receiving blocks of data can be seen as a multiple-input multiple-output (MIMO) communication system. Multiple data symbols are transmitted over the channel, another set of symbols are received, and finally the transmitted symbols are estimated from the information in the received symbols. The interest in MIMO systems increased dramatically a decade ago when it was discovered that multiple antennas at both the transmitter and the receiver can be used to transmit data very efficiently. For sufficiently rich scattering environments it was shown in [Tel95, FG98] that the increase in capacity by using antenna arrays is linear with the minimum number of transmit or receive antennas. This means that we can send much more data compared to single-antenna systems,

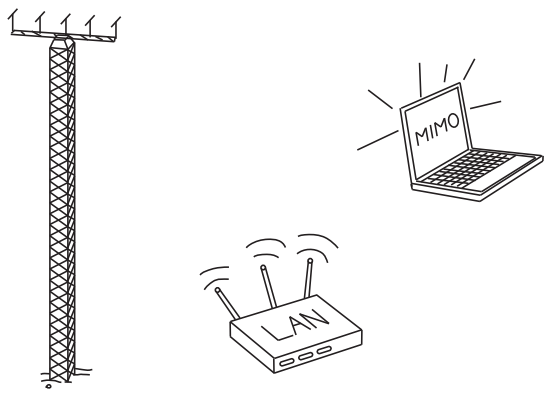

Figure 1.1: Multiple-antenna systems.

using the same amount of total power and the same amount of spectral resources. The idea is to multiplex data on parallel spatial subchannels, where the richness of the channel allows us to separate the subchannels on the receiver side. This is referred to as the multiplexing gain. Multiple antennas can also be used to provide a diversity gain in the MIMO channel, i.e. with more antennas the probability that all antennas experience a bad channel becomes smaller, making the system more robust to fading [TSC98, HH02]. Using multiple antennas to increase data rate through the multiplexing gain, or to make the system more robust through the diversity gain is a trade off [ZT03]. Recently, the use of multiple antennas at both the transmitter and the receiver has been specified in many wireless standards, such as the 3GPP LTE[1] standard [DPSB07]. However, as of today it is fair to say that the use MIMO systems has not yet reached its full potential in a commercial sense.

1.3 Delay-limited communication

One problem that comes with powerful channel coding is the delay it brings to the system. Wireless systems typically suffer from very unpredictable channels that change over time. For such cases, long error correcting codes can be problematic since the amount of *error protection* that is needed may change over the duration of a codeword. Retransmission of incorrectly decoded data is another common method for error protection in wireless communication systems. A checksum indicates to the receiver if a block of data has been incorrectly decoded, and if so, a request for a retransmission of the block is fed back to the transmitter. The longer the blocks, the more data has to be retransmitted once an error occurs, which further adds to the delay in the system.

[1]Third generation partnership project long term evolution.

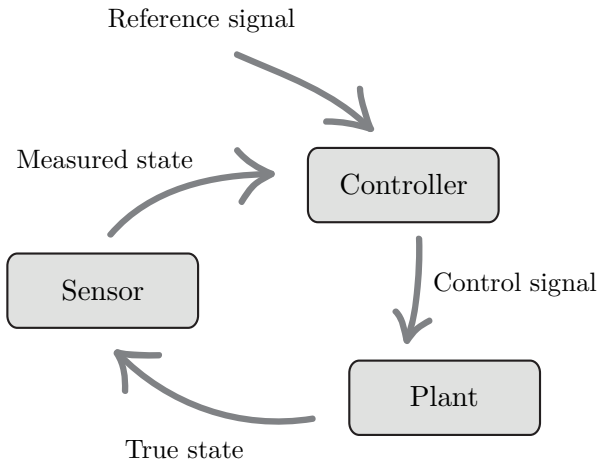

Figure 1.2: Closed-loop control system.

Some applications, such as closed-loop control applications are very sensitive to delay. A control system with a closed loop (see Figure 1.2) consists of a sensor (for example a radar) that measures some entity (the position of a rocket), a controller that sends control signals (rocket thrust), and a plant (a rocket). The state of the plant (the position) will subsequently affect the measurement, hence the term closed loop. Now, assume that parts of the control system are spatially dispersed and that communication between the entities has to be performed over radio. Then it is of outmost importance that this communication comes with as short delay as possible in order to preserve the stability of the system. A control application typically communicates at a fixed data rate (the control signaling is fixed), with the main objective to convey this information as quickly as possible, with a low power consumption and with a small probability of decoding errors. The above example motivates the analysis of a system with a fixed data rate and relatively short codewords that are delay limited.

In most cases it is mathematically intractable to provide a global performance analysis of the delay-sensitive system as a whole, let alone to jointly optimize the system. By only optimizing the lowest layer of the communication chain, our hope is that the overall performance also can be brought close to the optimum. This motivates the separate analysis of the modulation part of the physical layer — before we apply (possible) outer error-correcting codes. When not considering error correcting codes, the system will suffer from an inherent non-zero probability of detection error. Thus, not only must the optimal design trade off uncoded data rate against power usage, the design needs to consider the error probability as well.

This thesis considers the delay-limited communication problem, where the data rate and the codeword length are given, and where the task is to convey the data

using minimum amount of power, or with a minimum probability of decoding error.

1.4 Linear precoding and bit loading

Under the assumption that the transmitter knows the channel perfectly, the capacity-optimal strategy is to linearly orthogonalize the MIMO channel (using a precoder), and then convey the data over the non-interfering orthogonal subchannels. Each subchannel supports a specific data rate that is determined by the strength (signal to noise ratio) of the corresponding subchannel [FE91, GC97]. The procedure of optimizing the subchannel data rates is denoted bit loading. Figure 1.3 illustrates a MIMO communication system. Bits of data are multiplexed to separate subchannels, where each subchannel has an individual bit rate that is determined by the bit loading. The data in each subchannel is modulated to data symbols, then these symbols are mixed using a linear precoder to form a vector of transmit signals. The transmit signals are transmitted using multiple antennas, distorted by the radio channel, and received using multiple receive antennas. Finally, the transmitted data symbols are estimated (detected) using some type of detection algorithm at the receiver.

As was the case for capacity-optimal transmission, the linear precoder can be designed to make the effective subchannels orthogonal — to eliminate all inter-symbol interference between the subchannels. For the delay-limited case this orthogonalizing precoder is not necessarily optimal, sometimes it is better to allow the subchannels to interfere with each other [DDLW03, PCL03]. For the delay-limited case the jointly optimal design of the linear precoder and bit loading is still an open problem. We know that the optimal detector at the receiver is the maximum-likelihood (ML) detector, cf. [DGC03]. However, the optimal detector may suffer from high computational complexity as the number of dimensions grow [JO05]. Therefore, we often consider suboptimal detection algorithms, such as the zero-forcing receiver, the minimum mean squared error (MMSE) receiver [Pro01], or the decision feedback (DF) detector [BP79, WFGV98, GC01]. The design of the optimal transmitter depends on which detection algorithm that is used. Given the information about the channel, the transmitter needs to determine the bit loading on the subchannels, but also how the subchannels are to interfere each other through the linear precoder. Obviously, without taking the choice of receiver algorithm into account it becomes difficult to consider optimization of the transmitter. In this thesis we study the effects of bit loading and orthogonalization given different types of receiver structures.

1.5 Outline and contributions

This section gives the outline of the thesis, highlights the contributions, and provides references to the articles where the results where (or will be) presented. The main body of the thesis is separated into two parts, the first part considers the

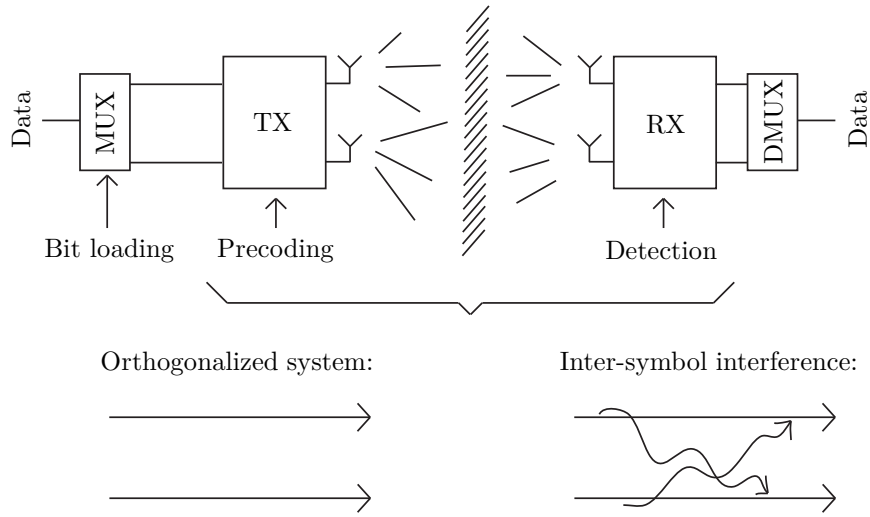

Figure 1.3: Bit loading and linear precoding.

bit loading and linear precoding problem assuming that the optimal ML detector is employed at the receiver, the second part considers the same problem but assuming the DF detector is used. Furthermore, the second part contains some extra mathematical results that are related to the DF design but will be treated separately in this outline.

Chapter 2: Background and problem formulation

Chapter 2 specifies and provides background to the MIMO communication problem, including mathematical model and assumptions. It contains the relevant references and some preliminary results that will serve as a basis for the discussion in the chapters that follow.

Chapters 3–5: Design based on maximum likelihood detection

The first part starts with Chapter 3, that introduces the design problem for ML detection. References to related work are provided. In Chapter 4, an approximation of the probability of detection error is derived that will serve as performance measure of the system, and the mathematical tools that are needed in order to optimize the transceiver are introduced. Chapter 5 presents the algorithm that optimizes the precoder and bit loading, numerical results and some analysis are presented.

We propose to use linear precoding and lattice invariant operations to transform the channel matrix into a lattice generator matrix with large minimum distance separation. With appropriate approximations, it is shown that this corresponds to

selecting lattices with good sphere-packing properties. Lattice invariant transformations are then used to minimize the power consumption. An algorithm for this power minimization is presented along with a lower bound on the optimization. Numerical results indicate significant gains by using the proposed method compared to channel diagonalization with adaptive bit loading.

The main contributions of Part I comprise of: The lattice based precoding algorithm that optimizes the transceiver, upper and lower bounds on the performance of the outcome of the algorithm, the motivation for using dense lattices in the context of linear precoding, and the observation that orthogonal subchannels are (in general) suboptimal given ML detection. The results in this part have previously been published in the following articles:

[**BO08**] S. Bergman and B. Ottersten. Lattice Based Linear Precoding for Multicarrier Block Codes. *IEEE Transactions on Signal Processing*, 56(7):2902–2914, July 2008.

[**BO07**] S. Bergman and B. Ottersten. Lattice Based Linear Precoding For MIMO Block Codes. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3:III329–III332, April 2007.

Chapters 6–9: Design based on decision feedback detection

The second part starts with Chapter 6 that provides a background to the transceiver design problem for DF detectors. In Chapter 7, a performance measure is derived and the design problem is formulated as a mathematical optimization problem. Chapter 8 considers the problem of designing the linear precoder together with the filters in the DF detector for a fixed bit loading. We show how this problem can be posed as a convex optimization problem, and we present an algorithm that solves this convex problem with linear complexity.

Chapter 9 considers the joint bit loading and precoder design problem. It is shown that the optimal design results in orthogonal subchannels, consequently we can apply conventional bit and power loading schemes for orthogonal subchannels to obtain the optimal transceiver. The proof is based on a relaxation of the discrete set of available bit rates on the individual subchannels to the set of positive real numbers. In practice, the signal constellations are discrete and the optimal relaxed bit loading has to be rounded. It is shown that the loss due to rounding is small, and an upper bound on the maximum loss is derived. Numerical results are presented that confirm the theoretical results and demonstrate that orthogonal transmission and the truly optimal DF design perform almost equally well.

The main contributions of Part II are: The observation that orthogonal subchannels are in fact optimal given DF detection, the algorithm that optimize the power allocation with linear complexity, the derivation of the optimal bit loading, and the discussion concerning the robustness with respect to rounding of the bit loading. The results in this part has previously been published in (or submitted as) the following articles:

[**BPO09**] S. Bergman, D.P. Palomar, and B. Ottersten. Optimal Bit Loading for MIMO Systems with Decision Feedback Detection. In *Proceedings IEEE Vehicular Technology Conference*, April 2009. Invited paper.

[**BPO08**] S. Bergman, D. P. Palomar, and B. Ottersten. Joint Bit Allocation and Precoding for MIMO Systems with Decision Feedback Detection. *IEEE Transactions on Signal Processing*, November 2008. Submitted.

Chapter 10: Skewed majorization

In the second part, Chapter 10, we present some related but rather self contained mathematical results regarding a class of optimization problems that arise in the precoder design for DF detectors. The class of problems is denoted optimization problems with skewed majorization constraints.

It is shown that the problem is equivalent to identifying the convex hull under a simple polygon defined by the parameters of the skewed majorization constraint. This leads to an algorithm that produces the exact optimum with linear computational complexity. As an application, we present two unitary precoder designs for a MIMO communication system with heterogenous signal constellations utilizing DF detection at the receiver. The results regarding skewed majorization constrained problems have previously been published in:

[**BJJO08**] S. Bergman, S. Järmyr, E. Jorswieck, and B. Ottersten. Optimization with Skewed Majorization Constraints: Application to MIMO Systems. In *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 1–6, September 2008.

Chapter 11: Thesis conclusions

This chapter concludes the thesis, and elaborates on possible lines for future research.

Notation

In this thesis, matrices are denoted by boldface, uppercase letters, $\boldsymbol{M}$, and vectors are denoted by boldface, lowercase letters, $\boldsymbol{v}$. Scalars are denoted by italic letters, e.g, x, K, α. The following mathematical notation will be used:

$\mathbb{C}^{N\times M}$	the set of complex-valued N by M matrices
$\mathbb{R}^{N\times M}$	the set of real-valued N by M matrices
$\mathbb{Z}^N$	the set of integer vectors of dimension N
$[\boldsymbol{M}]_{i,j}$	is the element on the i'th row and j'th column of $\boldsymbol{M}$
$\mathrm{E}[\cdot]$	statistical expectation
$\mathrm{vec}(\cdot)$	the vectorization operator on matrices
$\lvert\boldsymbol{M}\rvert$	the determinant of a matrix $\boldsymbol{M}$
$\mathrm{Tr}\{\boldsymbol{M}\}$	the trace of a matrix $\boldsymbol{M}$
$\lvert\boldsymbol{v}\rvert$	the Euclidian norm of a vector $\boldsymbol{v}$
$\boldsymbol{M}^{\mathrm{T}}$	the transpose of a matrix $\boldsymbol{M}$
$\boldsymbol{M}^*$	the conjugate transpose of a matrix $\boldsymbol{M}$
$\boldsymbol{M}^{-1}$	the inverse of a matrix $\boldsymbol{M}$
$\boldsymbol{d}(\boldsymbol{M})$	the vector of the diagonal elements of $\boldsymbol{M}$
$\boldsymbol{D}(\boldsymbol{m})$	the diagonal matrix with the diagonal elements $\boldsymbol{m}$
$\boldsymbol{D}(\boldsymbol{M})$	the diagonal matrix with the diagonal elements $\boldsymbol{d}(\boldsymbol{M})$
$N(\boldsymbol{m},\boldsymbol{R})$	the Gaussian multivariate distribution with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{R}$
$CN(\boldsymbol{m},\boldsymbol{R})$	the circularly symmetric complex Gaussian counterpart
$\Re\, c$	the real part of a complex number c
$\Im\, c$	the imaginary part of a complex number c
$\boldsymbol{X}\otimes\boldsymbol{Y}$	the Kronecker product of matrices, cf. [HJ91]
$\boldsymbol{I}_K$	the identity matrix of dimension K by K
$\boldsymbol{0}_{N\times M}$	the N by M matrix of only zeros
$\boldsymbol{1}_N$	the vector of all ones and length N by 1
$\boldsymbol{a}\preceq\boldsymbol{c}$	means $\boldsymbol{a}$ is majorized by $\boldsymbol{c}$, see e.g. Appendix 9.A, or [JB06]
$\boldsymbol{a}\preceq_\times\boldsymbol{c}$	means $\boldsymbol{a}$ is multiplicatively majorized by $\boldsymbol{c}$, see e.g. Appendix 9.A, or [JB06]
$\boldsymbol{a}\leq\boldsymbol{c}$	means $a_k\leq c_k$, for all vector indices k
$\nabla\mathcal{F}$	the gradient vector of a scalar-valued function $\mathcal{F}$
$O(N)$	the gradient vector of a scalar-valued function $\mathcal{F}$
$(x)^+$	maximum value of x and zero
$\lceil x\rfloor$	quantization of x
$\lVert\boldsymbol{x}\rVert_p$	the p-norm of a vector $\boldsymbol{x}$
arg max	the maximizing argument
arg min	the minimizing argument
$\triangleq$	defined as

Abbreviations

3GPP	third generation partnership project
AWGN	additive white Gaussian noise
CDF	cumulative distribution function
CR	cross (QAM constellation)
CSI	channel state information
BER	bit error rate
BLER	block error rate
BPSK	binary phase-shift keying
dB	decibel
DF	decision feedback
DFT	discrete fourier transform
GTD	generalized triangular decomposition
IID	independent identically distributed
KKT	Karush Kuhn Tucker (conditions)
LDPC	low-density parity-check (codes)
LTE	long term evolution
MAP	maximum a-posteriori
MIMO	multiple-input multiple-output
ML	maximum likelihood
MMSE	minimum mean squared error
MSE	mean squared error
PAM	pulse amplitude modulation
PEP	pairwise error probability
PSD	positive semi-definite
PSK	phase-shift keying
QAM	quadrature amplitude modulation
QR	(not an abbreviation, a matrix decomposition)
RX-CSI	receiver-side channel state information
SER	symbol error rate
SINR	signal to interference plus noise ratio
SISO	single-input single-output
SNR	signal to noise ratio
SVD	singular value decomposition
TH	Tomlinson-Harashima
TX-CSI	transmitter-side channel state information
ZF	zero forcing

Appendix 1.A Work not covered by the thesis

Some of our published work did not fit into the scope of this thesis; in this appendix we list them.

A transmitter can not estimate the channel while simultaneously transmitting on the same time–frequency slot. Therefore, it may not be reasonable to assume perfect knowledge about the channel at the transmitter during transmission. During my Ph.D. studies we have presented several different approaches to the transceiver design with partial or imperfect channel state information. The designs differ in the types of channel state information that are available: In particular either first order statistics, second order statistics, or both first and second order statistics where considered.

In [JBO08] we proposed a precoding scheme for the case of second order statistics given a DF detector at the receiver. In [MBO04b] we proposed a precoding scheme for the ML detector, also assuming second order statistics. In [MBO04a, BO06, BO05a, BO05b] we proposed different transmission schemes for ML detection given first order statistics of the channel, and, in [BMO04] we proposed a scheme for ML that can be used for both first and second order statistics.

When perfect channel state information is available, data can be multiplexed and optimized over independent spatial channels. When the available channel information is imperfect or partial, parallel orthogonal transmission is impossible and crosstalk between the spatial channels will inevitably complicate the analysis as well as the design. Due to the difficulties analyzing the system, our precoders based on imperfect or partial channel estimates can not easily be compared in closed form or analytically. For this reason we have decided not to include these results in the thesis.

[BO06] S. Bergman and B. Ottersten. Design of robust linear dispersion codes based on imperfect CSI for ML receivers. In *Proceedings European Signal Processing Conference*, September 2006.

[BO05a] S. Bergman and B. Ottersten. Adaptive spatial bit loading using imperfect channel state information. In *Proceedings of International Workshop on Optical and Electronic Device Technology for Access Networks, Aalborg, Denmark*, September 2005. Invited Paper.

[BO05b] S. Bergman and B. Ottersten. Spatial multiplexing over Rician fading channels: Linear precoding transmission strategies *Nordic Conference on Radio Science and Communications (RVK)*, June 2005.

[BMO04] S. Bergman, C. Martin, and B. Ottersten. Bit and Power Loading for Spatial Multiplexing using Partial Channel State Information. In *Proceedings ITG Workshop on Smart Antennas, Technische Universität Munich*, 152–159, March 2004.

[**JBO08**] S. Järmyr, S. Bergman, and B. Ottersten. Long-Term Adaptive Precoding for Decision Feedback Equalization. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2897–2900, April 2008.

[**MBO04b**] C. Martin, S. Bergman, and B. Ottersten. Spatial loading based on channel covariance feedback and channel estimates. In *Proceedings European Signal Processing Conference*, 519–522, September 2004.

[**MBO04a**] C. Martin, S. Bergman, and B. Ottersten. Simple Spatial Multiplexing Based on Imperfect Channel Estimates. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 713–716, May 2004.

Chapter 2

Background and problem formulation

In this chapter we introduce the system model, provide the relevant background, present the main assumptions, and define the problem that will be considered in the later chapters.

2.1 System model

Consider the discrete-time linear model of an $N_r \times N_t$ MIMO baseband symbol-sampled communication system over the set of complex numbers $\mathbb{C}$. Such a system can be modeled using the linear regression equation

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}, \tag{2.1}$$

where $\boldsymbol{y} \in \mathbb{C}^{N_r}$ is the received signal, $\boldsymbol{H} \in \mathbb{C}^{N_r \times N_t}$ is the channel matrix, $\boldsymbol{x} \in \mathbb{C}^{N_t}$ is the vector of transmitted signals, and $\boldsymbol{n} \in \mathbb{C}^{N_r}$ is a vector with additive white circularly-symmetric complex-Gaussian noise. Further, the noise vector has covariance matrix $\mathrm{E}\{\boldsymbol{n}\boldsymbol{n}^*\} = \boldsymbol{R}_{\boldsymbol{n}}$.

In some cases it is easier to work in the real-valued rather than the complex-valued domain. The complex-valued system equation can be reformulated to a real-valued equation using

$$\boldsymbol{y}_r = \begin{bmatrix} \Re\boldsymbol{y} \\ \Im\boldsymbol{y} \end{bmatrix}, \ \boldsymbol{x}_r = \begin{bmatrix} \Re\boldsymbol{x} \\ \Im\boldsymbol{x} \end{bmatrix}, \ \boldsymbol{n}_r = \begin{bmatrix} \Re\boldsymbol{n} \\ \Im\boldsymbol{n} \end{bmatrix}, \tag{2.2}$$

$$\boldsymbol{H}_r = \begin{bmatrix} \Re\boldsymbol{H} & \Im\boldsymbol{H} \\ -\Im\boldsymbol{H} & \Re\boldsymbol{H} \end{bmatrix}, \tag{2.3}$$

where $\Re$ extracts the real part, and $\Im$ the imaginary part of the argument. It is straightforward to verify that the real-valued system model

$$\boldsymbol{y}_r = \boldsymbol{H}_r\boldsymbol{x}_r + \boldsymbol{n}_r, \tag{2.4}$$

is equivalent to (2.1) with the noise vector, $\tilde{\boldsymbol{n}}_r$, Gaussian distributed as

$$\boldsymbol{n}_r \sim N\left(\boldsymbol{0}, \frac{1}{2}\begin{bmatrix} \Re\boldsymbol{R}_n & -\Im\boldsymbol{R}_n \\ \Im\boldsymbol{R}_n & \Re\boldsymbol{R}_n \end{bmatrix}\right). \tag{2.5}$$

In this thesis, both the complex-valued and the real-valued equations will be used to describe the MIMO system.

2.1.1 Channel state information

The receiver can typically estimate the channel from the received signal with the help of training sequences, or piloting symbols. If the channel is sufficiently stationary, a receiver-side channel estimate obtained using sufficient amount of training can be treated as an exact description of the true channel matrix. We say that the receiver-side channel state information (RX-CSI) is perfect. On the transmitter side, one can not directly estimate the channel since it is not possible to receive and transmit on the same frequency and time-slot. Indirect methods for the transmitter to obtain the channel estimate include

- *Obtaining the information from the receiver using a feedback link.* A drawback with this method is the bandwidth resources that are consumed by the feedback link. There is also an inherent delay in the system that can make the channel information outdated when it becomes available to the transmitter.
- *Using the reciprocity property of the channel*, i.e., using the fact that the forward channel is equivalent to the reverse channel. Problems with this approach includes issues with calibration and that the forward and reverse links are not necessarily close in frequency and time.

Despite the practical difficulties to obtain perfect transmitter-side channel state information (TX-CSI), if there is a two-way communication with sufficient capacity and if the channel varies slowly, then it is possible to assume perfect TX-CSI. Unless explicitly stated otherwise, in this thesis we make the assumption that both the transmitter and the receiver know the channel matrix $\boldsymbol{H}$ and the noise covariance matrix $\boldsymbol{R_n}$ perfectly.

2.1.2 Applications

Up to this point, the origin of the N_t input and N_r output signals has not been specified. The input and output signals can be obtained from various sources, such as different samples in time, frequency, multiple antennas, or any combination of the three. In the case when there are uncertainties in the CSI, the channel statistics are often modeled based on how the input and output vectors where collected. Herein, where the channel matrix is known exactly, the origin of the vectors are of minor importance to the transmitter and the receiver. However, in order to illustrate how (2.1) can be used in practice we will consider two examples.

In the first example, we consider a MIMO system with N_t transmitting antennas, N_r receiving antennas, and N_c orthogonal frequencies (sub-carriers) that have been orthogonalized using orthogonal frequency-division multiplexing (OFDM). Each sub-carrier can be modeled as

$$\boldsymbol{y}_n = \boldsymbol{H}_n \boldsymbol{x}_n + \boldsymbol{n}_n, \qquad \forall \quad n = 0, ..., N_c - 1, \tag{2.6}$$

where n denotes the sub-carrier index. For each sub-carrier the vector $\boldsymbol{y}_n \in \mathbb{C}^{N_r}$ denotes received signals, $\boldsymbol{x}_n \in \mathbb{C}^{N_t}$ is the vector of transmitted signals, $\boldsymbol{H}_n \in \mathbb{C}^{N_r \times N_t}$ is the channel matrix, and finally $\boldsymbol{n}_n \in \mathbb{C}^{N_r}$ is the noise vector. The noise is assumed to be independent identically-distributed (IID) circularly-symmetric complex Gaussian, zero-mean with variance one for each component. By stacking the equations of the sub-carriers

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_0 \\ \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_{L-1} \end{bmatrix}, \ \boldsymbol{n} = \begin{bmatrix} \boldsymbol{n}_0 \\ \boldsymbol{n}_1 \\ \vdots \\ \boldsymbol{n}_{L-1} \end{bmatrix}, \ \boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_0 \\ \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_{L-1} \end{bmatrix}, \tag{2.7}$$

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_0 & & \\ & \ddots & \\ & & \boldsymbol{H}_{Q-1} \end{bmatrix}, \tag{2.8}$$

we obtain the system model as in (2.1).

In the second example, we model a finite impulse response channel with colored additive noise for a wireless communication system with one transmitting and one receiving antenna. Assume that the discrete-time impulse response of the channel can be approximated using a finite number of taps, $h_0, ..., h_{N_h-1}$. Also, assume that the transmitter sends a complex-valued codeword $\boldsymbol{x} \in \mathbb{C}^{N_l}$. Let the receiver collect the corresponding received symbols in a vector $\boldsymbol{y} \in \mathbb{C}^{N_h+N_l-1}$. The received signal may be subject to some colored additive, zero-mean complex-Gaussian noise, distributed as $\boldsymbol{n} \sim CN(\boldsymbol{0}_{N_h+N_l-1}, \boldsymbol{R})$. Furthermore, neglecting the inter-block interference (for example by applying zero-padding), the system can be modeled as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}, \tag{2.9}$$

where the channel matrix is formed as

$$\boldsymbol{H} = \begin{bmatrix} h_0 & & \\ \vdots & \ddots & \\ h_{N_h-1} & & h_0 \\ & \ddots & \vdots \\ & & h_{N_h-1} \end{bmatrix}. \tag{2.10}$$

The above examples show the flexibility of the MIMO system. For some applications it can be beneficial to utilize specific structures in $\boldsymbol{H}$ to reduce the computational complexity of the system. In this thesis however, in order not to lose generality, we will not make such structural assumptions and $\boldsymbol{H}$ can be of arbitrary structure.

2.2 The MIMO communication problem

With the system model in place, our next step is to define the communication problem. Assume that both the transmitter and the receiver are aware of a codebook, $\mathcal{X} \subset \mathbb{C}^{N_t}$, consisting of a set of transmit vectors, $\boldsymbol{x} \in \mathcal{X}$. Each vector represent a unique number (a sequence of bits) that can be sent from the transmitter to the receiver. Based on the information that is to be communicated[1], the transmitter picks a transmit vector from the codebook, then sends it over the channel (2.1) to the receiver. The receiver decodes the transmitted bits by estimating the vector $\boldsymbol{x}$ from the received vector $\boldsymbol{y}$, using its knowledge about the channel and the codebook as side information.

The problem considered in this thesis is to design the codebook of transmit vectors, $\mathcal{X}$, to convey bits of information over the MIMO channel under an average transmit power constraint

$$\mathrm{E}\{\boldsymbol{x}^*\boldsymbol{x}\} \leq P. \tag{2.11}$$

It is assumed that the CSI is perfect; both the transmitter and the receiver knows the channel matrix $\boldsymbol{H}$ and the noise covariance $\boldsymbol{R_n}$ perfectly.

Note that at this point we do not specify what the objective of the design is. An objective could be, for example, to maximize the mutual information between $\boldsymbol{y}$ and $\boldsymbol{x}$, or to minimize the probability of detection error given a certain codebook size. Before we look at specific problems, we will first consider two objective-independent procedures that can make the problem easier to handle; the noise pre-whitening, and the parallel single-input single-output transmission (SISO).

2.2.1 Noise pre-whitening

Because a rank-deficient noise covariance matrix $\boldsymbol{R_n}$ would imply infinite signal to interference plus noise ratio (SINR), we can assume that any valid communication problem has a non-singular $\boldsymbol{R_n}$. Multiplying the received signal $\boldsymbol{y}$ with a non-singular matrix does not remove any information about the transmitted signal $\boldsymbol{x}$ from $\boldsymbol{y}$. Therefore, we can perform noise whitening of the received signal according to

$$\boldsymbol{y}_w \triangleq \boldsymbol{R_n}^{-\frac{1}{2}}\boldsymbol{y} = \boldsymbol{R_n}^{-\frac{1}{2}}\boldsymbol{H}\boldsymbol{x} + \boldsymbol{R_n}^{-\frac{1}{2}}\boldsymbol{n} \triangleq \boldsymbol{H}_w\boldsymbol{x} + \boldsymbol{n}_w, \tag{2.12}$$

where $\boldsymbol{y}_w$, $\boldsymbol{H}_w$, and $\boldsymbol{n}_w$ are the pre-whitened counterparts of $\boldsymbol{y}$, $\boldsymbol{H}$, and $\boldsymbol{n}$, respectively. The main difference is, of course, that the pre-whitened noise is white

[1] Assume here that the information is sufficiently random for us to consider all transmit vectors in $\mathcal{X}$ as equally likely.

$\mathrm{E}\{\boldsymbol{n}_w \boldsymbol{n}_w^*\} = \boldsymbol{I}$. In this work, from this point and onwards, we will assume that the noise is uncorrelated, or similarly, that pre-whitening has already been performed. Whenever the original system model (2.1) is referred to, it is assumed implicitly that $\boldsymbol{R}_{\boldsymbol{n}} = \boldsymbol{I}$. This implies no loss of generality.

2.2.2 Equivalent parallel SISO system

If the transmitter and the receiver can perform linear filtering of the transmitted and received signals, then the MIMO channel can be transformed into several parallel SISO subchannels. Introduce the singular value decomposition of the channel matrix

$$\boldsymbol{H} = \boldsymbol{U}_{\boldsymbol{H}} \boldsymbol{\Lambda}_{\boldsymbol{H}} \boldsymbol{V}_{\boldsymbol{H}}^*, \tag{2.13}$$

where $\boldsymbol{\Lambda}_{\boldsymbol{H}}$ is a diagonal matrix (with real-valued decreasing elements on the diagonal) and $\boldsymbol{U}_{\boldsymbol{H}}$, $\boldsymbol{V}_{\boldsymbol{H}}$ are unitary matrices. Since unitary matrices are non-singular, no information is lost by multiplying the received and transmitted signal vectors with unitary matrices as

$$\tilde{\boldsymbol{y}} = \boldsymbol{U}_{\boldsymbol{H}}^* \boldsymbol{y}, \quad \tilde{\boldsymbol{x}} = \boldsymbol{V}_{\boldsymbol{H}}^* \boldsymbol{x}, \quad \tilde{\boldsymbol{n}} = \boldsymbol{U}_{\boldsymbol{H}}^* \boldsymbol{n}. \tag{2.14}$$

Note that the power constraint $\mathrm{E}\{\tilde{\boldsymbol{x}}^* \tilde{\boldsymbol{x}}\} \leq P$ is preserved, and the covariance matrix of the noise vector is $\mathrm{E}\{\tilde{\boldsymbol{n}} \tilde{\boldsymbol{n}}^*\} = \boldsymbol{I}$. Actually, since the noise is complex Gaussian zero-mean, its distribution is also preserved with the matrix rotation. The linearly pre-filtered system model is then

$$\tilde{\boldsymbol{y}} = \boldsymbol{\Lambda}_{\boldsymbol{H}} \tilde{\boldsymbol{x}} + \tilde{\boldsymbol{n}}, \tag{2.15}$$

which corresponds to parallel SISO channels. Hence, any MIMO system on the form (2.1) with perfect CSI at the transmitter as well as the receiver can be transformed to (2.15) without loss of generality.

2.3 Capacity-optimal transmission

From information theory we know that the highest rate at which information can be conveyed over an additive white Gaussian noise (AWGN) channel is given by the maximum mutual information between the transmitted and the received signals (cf. [CT91]). Here, we recapitulate how to obtain the solution to the maximum mutual information problem [CT91]. The resulting codebook attains the capacity of the MIMO channel, and therefore we refer to this transmission scheme as the capacity-optimal transmission.

The following lemma states that the equivalent parallel SISO system with independent subchannels can maximize the mutual information of the MIMO system.

Lemma 2.3.1 *It is optimal in terms of maximum mutual information to send data independently over parallel orthogonal subchannels.*

Proof: The mutual information between the received signal, $\tilde{\boldsymbol{y}}$, and the transmitted signal, $\tilde{\boldsymbol{x}}$, is given by

$$\mathcal{I}(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}}) = h(\tilde{\boldsymbol{y}}) - h(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}}) = h(\tilde{\boldsymbol{y}}) - h(\tilde{\boldsymbol{n}}), \tag{2.16}$$

where $h(\tilde{\boldsymbol{y}})$ and $h(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{x}})$ denotes entropy and conditional entropy, respectively. Since the elements of $\tilde{\boldsymbol{n}}$ are statistically independent we have

$$h(\tilde{\boldsymbol{n}}) = \sum_{i=1}^{N_r} h(\tilde{n}_i), \tag{2.17}$$

and using that the sum entropy is larger than or equal to the joint entropy, we get

$$\mathcal{I}(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}}) \leq \sum_{i=1}^{N_r} h(\tilde{y}_i) - h(\tilde{n}_i) = \sum_{i=1}^{N_t} h(\tilde{y}_i) - h(\tilde{n}_i). \tag{2.18}$$

The inequality is satisfied with equality if the elements of $\tilde{\boldsymbol{x}}$ are mutually independent. □

Maximizing the mutual information over an AWGN SISO channel results in Gaussian distributed codebooks [Sha48]. Using Lemma 2.3.1, the mutual information of a MIMO channel is therefore maximized if $\tilde{\boldsymbol{x}}$ satisfies

$$\tilde{\boldsymbol{x}} \sim CN\big(\boldsymbol{0}, \boldsymbol{P}\big), \tag{2.19}$$

where $\boldsymbol{P}$ is diagonal and positive (so that the symbols are independent). The mutual information is then given by

$$\mathcal{I}(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}}) = \sum_{i=1}^{N_t} \log(1 + \lambda_i p_i), \tag{2.20}$$

where the channel gain of subchannel i is denoted $\lambda_i = [\boldsymbol{\Lambda}_H^2]_{i,i}$, and $p_i = [\boldsymbol{P}]_{i,i}$ is the corresponding transmit power. The constraints on the power allocation are

$$\sum_{i=1}^{N_t} p_i \leq P, \quad p_i \geq 0 \quad \forall\, i = 1, ..., N_t. \tag{2.21}$$

Maximizing the mutual information (2.20) with respect to the power under the above constraints is a convex problem. The solution is given by the so-called water-filling solution [CT91]

$$p_i = (\mu - \lambda_i^{-1})^+, \qquad \forall\, i = 1, ..., N_t, \tag{2.22}$$

where the water level, μ, is chosen such that the sum-power constraint is satisfied with equality

$$\sum_{i=1}^{N_t} p_i = P. \tag{2.23}$$

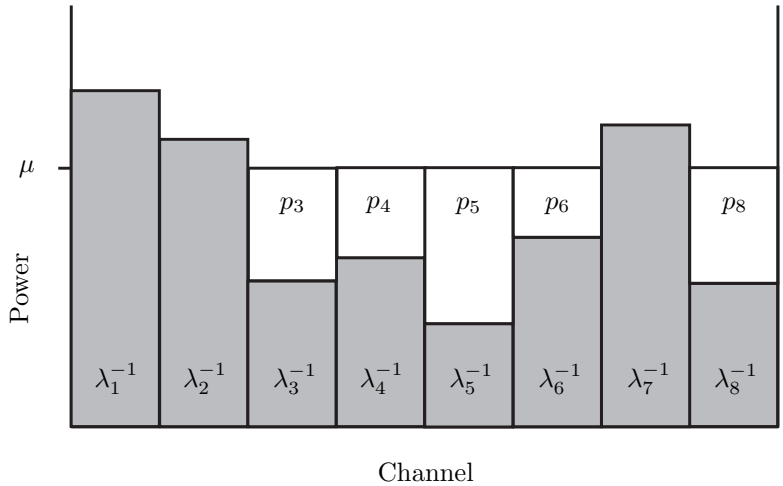

Figure 2.1: Illustration of the water-filling power allocation.

The analogy of water filling is illustrated in Figure 2.1: The power allocation can be seen as the depth of water that has been poured on a 'seabed' represented by the noise powers $\lambda_1^{-1}, ..., \lambda_{N_t}^{-1}$. Strong subchannels with a low noise power get more power than weak subchannels. The water level, μ, determines the number of active subchannels. By pouring more water (i.e. by increasing the total transmit power), more subchannels are activated as their corresponding noise powers are submerged.

2.3.1 Practical considerations

Gaussian distributed codebooks make the decoding procedure very difficult due to the lack of structure in the code. In order to attain the capacity it is necessary to repeatedly use the channel a large number of times (in fact an infinite number of times), and then jointly detect the block of transmitted vectors as one big codeword. The detector has to search through all possible codeword blocks in the infinite codebook, which is not feasible in practice. In addition to the computational complexity of the search, the channel may change over time and therefore our assumption of perfect CSI becomes difficult to attain.

However, the capacity optimal transmission can serve as an upper bound or a benchmarking scheme to other more practical schemes that strive to maximize the data rate with vanishing probability of error. By using state of the art error correcting codes, such as low-density parity-check (LDPC) codes (cf. [RU08]), it is possible to convey data at a rate very close to the capacity. These LDPC codes still need to be of a rather high dimensionality ($\sim 10^5$) in order to approach the capacity, see e.g., [LYW04, tBKA04, BB06]. The high dimensionality of the codeword block introduces a delay in the system that may be problematic for certain types of

applications. This motivates the next discussion on delay-limited transmission.

2.4 Delay-limited transmission

One downside with the capacity-optimal transmission scheme is the (infinitely) long codeword blocks and the delay that this brings to the system. Clearly, a long delay has practical disadvantages; especially when considering time-varying channels, systems with packet retransmission, or delay-sensitive applications. As an example of a delay-sensitive application, we can consider the control system that was discussed in Chapter 1. Ideally such a system needs to be reliable (low error probability), power efficient, but perhaps most importantly — it needs to have a short delay. In this case, achieving a high data rate is of minor importance if it comes at a cost of instability due to delay.

Without error correcting codes the system will suffer from an inherent non-zero probability of detection error. Thus, not only must the optimal design trade off uncoded data rate against power usage; it needs to consider the probability of detection error as well. We define the delay-limited communication problem, which will be the main focus of this thesis, as follows: The objective is to convey a certain number of bits, R, of data over the channel (2.1), under an average power constraint

$$\mathrm{E}\{\boldsymbol{x}^*\boldsymbol{x}\} \leq P, \tag{2.24}$$

and with minimum probability of detection error. It is assumed that the RX-CSI as well as the TX-CSI is perfect. At this point we have not defined how the receiver detects the transmit vector, and thus it is not clear what the probability of detection error is. In Section 2.7, we specify the detection algorithms that will be considered in the thesis.

Note how this problem formulation differs from the capacity-optimal transmission, where the focus is on transmitting at a certain bit rate as opposed to transmitting a fixed number of bits. To see the difference, consider a case when we transmit R bits, using vectors of dimension N_t, and with a transmit power P. By instead concatenating L transmit vectors to one vector, we transmit $L\,R$ bits using a vector of dimension $L\,N_t$ and with a total transmit power $L\,P$. Even though the data rate and average power consumption per dimension remains unchanged with the concatenated vector, the delay is not the same and consequently the problems are not comparable in the delay-limited sense. If we instead focus on average data throughput, both cases satisfy the constraints on rate and power, and since the probability of error goes down with increasing L, the capacity-optimal solution is to use infinitely long codewords.

Now that the delay-limited problem has been formulated, we continue with the design of a codebook, $\mathcal{X}$. Optimizing a codebook without any imposed structure on the codewords is very difficult. One way to introduce structure is by means of linear precoding which is the topic of the following section.

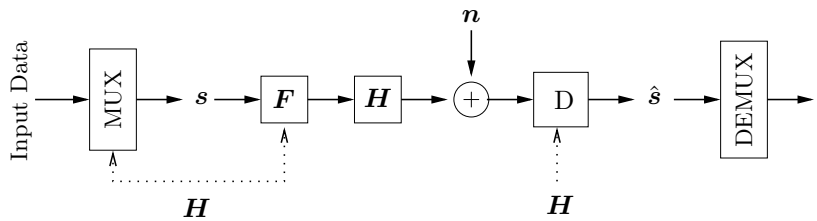

Figure 2.2: The linearly precoded system. Data is multiplexed and modulated to form a symbol vector $\boldsymbol{s}$. The vector is linearly precoded using $\boldsymbol{F}$, sent over the linear channel $\boldsymbol{H}$ with AWGN $\boldsymbol{n}$, and then detected on the receiver side using a detection algorithm D. The data is finally extracted from the detected symbol vector $\hat{\boldsymbol{s}}$.

2.5 Linear precoding

In Section 2.3 it was shown that the mutual information can be maximized by using correlated complex-Gaussian distributed transmit vectors. The transmit vectors are constructed as

$$\boldsymbol{x} = \boldsymbol{F}\boldsymbol{s}, \tag{2.25}$$

where the matrix $\boldsymbol{F} \in \mathbb{C}^{N_t \times N}$ is a data-independent correlating matrix (that will be referred to as the precoder), and where $\boldsymbol{s} \in \mathbb{C}^N$ is a vector containing the data symbols that are drawn from the complex-Gaussian distribution as

$$\boldsymbol{s} \sim CN(\boldsymbol{0}, \boldsymbol{I}). \tag{2.26}$$

For capacity-optimal transmission, the precoder should be chosen as

$$\boldsymbol{F} = \boldsymbol{V}_{\boldsymbol{H}} \boldsymbol{P}^{1/2}, \tag{2.27}$$

where the diagonal matrix $\boldsymbol{P}$ satisfies the water-filling equations (2.22) and (2.23). Given that the receiver multiplies the received signal vector with the unitary matrix $\boldsymbol{U}_{\boldsymbol{H}}^*$, this precoder creates orthogonal, parallel SISO channels. Interestingly, this is not the only optimal precoder; any precoder on the form

$$\boldsymbol{F} = \boldsymbol{V}_{\boldsymbol{H}} \boldsymbol{P}^{1/2} \boldsymbol{Q}^*, \tag{2.28}$$

where $\boldsymbol{Q}$ is unitary is also optimal, since the complex-Gaussian vectors $\boldsymbol{s}$ and $\boldsymbol{Q}\,\boldsymbol{s}$ have the same distribution. We say that the transmit vector, $\boldsymbol{x}$, is obtained using a linear precoding, $\boldsymbol{F}$, of a data symbols vector, $\boldsymbol{s}$.

This structure can be generalized to non-Gaussian data symbols. We define linear precoding as the synthesis of the transmit signal $\boldsymbol{x}$, using a linear combination of independent random variables (not necessarily Gaussian) stacked in a symbol vector, $\boldsymbol{s}$, as

$$\boldsymbol{x} = \boldsymbol{F}\boldsymbol{s}. \tag{2.29}$$

The matrix $\boldsymbol{F}$ is the data independent precoding matrix, also denoted the precoder. Without loss of generality, the symbol vector is normalized as

$$\mathrm{E}\{\boldsymbol{s}\boldsymbol{s}^*\} = \boldsymbol{I}, \tag{2.30}$$

which implies that the power constraint (2.11) becomes a function of the precoder as

$$\mathrm{Tr}\{\boldsymbol{F}\boldsymbol{F}^*\} \leq P. \tag{2.31}$$

The linearly precoded system is illustrated in Figure 2.2. It will be of interest to consider the following SVD-like decomposition of the precoder

$$\boldsymbol{F} = \boldsymbol{U}_{\boldsymbol{F}}\boldsymbol{\Sigma}_{\boldsymbol{F}}\boldsymbol{Q}^*, \tag{2.32}$$

where $\boldsymbol{U}_{\boldsymbol{F}} \in \mathbb{C}^{N_t \times N}$ has orthonormal columns and determines the directivity of the precoder, $\boldsymbol{\Sigma}_{\boldsymbol{F}} = \boldsymbol{P}^{1/2}$ is diagonal and specifies the power assigned to the spatial subchannels, and finally $\boldsymbol{Q}$ is unitary and determines how the symbol vector is mixed (or rotated) before power allocation. For reasons that will be explained in the next subsection we do not yet impose a specific ordering of the diagonal elements of $\boldsymbol{\Sigma}_{\boldsymbol{F}}$.

Although linear precoding is (in general) suboptimal for delay-limited transmission, it remains a very attractive transmission strategy due to its simplicity; it is straightforward to implement, and easy to adapt to various channel conditions.

2.5.1 Optimal directivity matrix

The following lemma shows the optimal transmit directivity matrix of the linearly precoded system. The lemma is based on well known results from matrix analysis [HJ85], and this result (or similar variants) occurs frequently in the MIMO literature, cf. [Tel95, PCL03, SD07, JSBC04, KS04, HJ85, SD08]. The lemma is central in the linear precoding design, and therefore we will give our version of the proof in full detail.

Lemma 2.5.1 *The power-optimal linear precoder transmits in the directions of the eigenvectors of the channel matrix, and assigns power to the eigenmodes of the channel such that the order of the eigenvalues of the effective channel remains unchanged.*

Proof: The system model

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{F}\boldsymbol{s} + \boldsymbol{n}, \tag{2.33}$$

can be reformulated to the following equivalent parallel SISO form

$$\boldsymbol{y}_{\mathrm{AWGN}} = \boldsymbol{s} + \boldsymbol{n}_{\mathrm{AWGN}}, \tag{2.34}$$

where the random variables in $\boldsymbol{s}$ are statistically independent of $\boldsymbol{F}$, and where the noise distribution is given by

$$\boldsymbol{n}_{\mathrm{AWGN}} \sim CN\big(\boldsymbol{0}, (\boldsymbol{F}^*\boldsymbol{H}^*\boldsymbol{H}\boldsymbol{F})^{-1}\big). \tag{2.35}$$

By keeping $\boldsymbol{F}^*\boldsymbol{H}^*\boldsymbol{H}\boldsymbol{F} \in \mathbb{C}^{N\times N}$ fixed, the performance (in terms of detection error) of the system becomes independent of $\boldsymbol{F}$. Denote the eigenvector decomposition of $\boldsymbol{H}^*\boldsymbol{H} = \boldsymbol{V}_{\boldsymbol{H}}\boldsymbol{\Lambda}_{\boldsymbol{H}}^2\boldsymbol{V}_{\boldsymbol{H}}^*$, where $\boldsymbol{V}_{\boldsymbol{H}} \in \mathbb{C}^{N_r\times Q}$ contains orthonormal columns and $\boldsymbol{\Lambda}_{\boldsymbol{H}} \in \mathbb{R}^{Q\times Q}$ is diagonal and positive definite. Further, define the matrix

$$\boldsymbol{A} = \boldsymbol{\Lambda}_{\boldsymbol{H}}\boldsymbol{V}_{\boldsymbol{H}}^*\boldsymbol{F} \in \mathbb{C}^{Q\times N}. \tag{2.36}$$

Note that, by assumption, $\boldsymbol{A}^*\boldsymbol{A}$ is a fixed matrix. Note also that $Q \geq N$. The transmitted power is

$$\begin{aligned}\mathrm{Tr}\{\boldsymbol{F}\boldsymbol{F}^*\} &= \mathrm{Tr}\{\boldsymbol{\Lambda}_{\boldsymbol{H}}^{-2}\boldsymbol{A}\boldsymbol{A}^*\} \\ &= \boldsymbol{d}\big(\boldsymbol{\Lambda}_{\boldsymbol{H}}^{-2}\big)^{\mathrm{T}}\, \boldsymbol{M}\, \boldsymbol{\sigma}(\boldsymbol{A}\boldsymbol{A}^*),\end{aligned} \tag{2.37}$$

where $\boldsymbol{\sigma}(\boldsymbol{A}\boldsymbol{A}^*)$ are the (real non-negative) eigenvalues of $\boldsymbol{A}\boldsymbol{A}^*$ sorted in decreasing order, and $\boldsymbol{M}$ is a doubly stochastic matrix. Because $Q \geq N$, we have the following relation

$$\boldsymbol{\sigma}(\boldsymbol{A}\boldsymbol{A}^*) = [\boldsymbol{\sigma}(\boldsymbol{A}^*\boldsymbol{A})^{\mathrm{T}}\ \boldsymbol{0}_{1\times(Q-N)}]^{\mathrm{T}}, \tag{2.38}$$

and because $\boldsymbol{A}^*\boldsymbol{A}$ is fixed, we therefore know that $\boldsymbol{\sigma}(\boldsymbol{A}\boldsymbol{A}^*)$ is fixed. The stochastic matrix, $\boldsymbol{M}$, is thus the only parameter that depends on $\boldsymbol{F}$ given that $\boldsymbol{A}^*\boldsymbol{A}$ is fixed. Hence, our problem is to find the minimizing $\boldsymbol{M}$ in the set of doubly stochastic matrices. Any doubly stochastic matrix can be written as a convex combination of all permutation matrices of the same dimension [HJ85]. Let $\boldsymbol{\Pi}_{(1)}, ..., \boldsymbol{\Pi}_{(Q!)}$ be the enumeration of all permutation matrices, then

$$\mathrm{Tr}\{\boldsymbol{F}^*\boldsymbol{F}\} \geq \min_i \boldsymbol{d}(\boldsymbol{\Lambda}_{\boldsymbol{H}}^{-2})^{\mathrm{T}}\, \boldsymbol{\Pi}_{(i)}\, \boldsymbol{\sigma}(\boldsymbol{A}\boldsymbol{A}^*). \tag{2.39}$$

Clearly, if the eigenvalues, $\boldsymbol{\Lambda}_{\boldsymbol{H}}$, and $\boldsymbol{\sigma}(\boldsymbol{A}^*\boldsymbol{A})$ are ordered in decreasing order, the minimizing permutation matrix is the identity matrix $\boldsymbol{\Pi}_{(i)} = \boldsymbol{I}$. The lower bound is attained when the precoder is on the form

$$\boldsymbol{F} = \boldsymbol{V}_{\boldsymbol{H}} \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{F}} \\ \boldsymbol{0}_{(Q-N)\times N} \end{bmatrix} \boldsymbol{Q}^*, \tag{2.40}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{F}}$ is non-negative diagonal such that $\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H}}\boldsymbol{\Sigma}_{\boldsymbol{F}}$ is decreasing along the diagonal, where $\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H}}$ is the upper-left $N \times N$ block of $\boldsymbol{\Lambda}_{\boldsymbol{H}}$. The right-side unitary matrix, $\boldsymbol{Q}$, of the precoder has the following impact on the fixed matrix

$$\boldsymbol{F}^*\boldsymbol{H}^*\boldsymbol{H}\boldsymbol{F} = \boldsymbol{Q}\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H}}^2\boldsymbol{\Sigma}_{\boldsymbol{F}}^2\boldsymbol{Q}^*. \tag{2.41}$$

□

Now, by minimizing the transmitted power $\mathrm{Tr}\{\boldsymbol{F}\boldsymbol{F}^*\}$ subject to a fixed $\boldsymbol{F}^*\boldsymbol{H}^*\boldsymbol{H}\boldsymbol{F}$, we obtain a Pareto optimum that also must be satisfied for the system using full power, $\mathrm{Tr}\{\boldsymbol{F}\boldsymbol{F}^*\} = P$, with a rescaled $\boldsymbol{F}^*\boldsymbol{H}^*\boldsymbol{H}\boldsymbol{F}$.

Note the differences and similarities between Lemmas 2.3.1 and 2.5.1, where the former shows that in order to maximize the mutual information it is necessary to transmit linearly precoded Gaussian symbols, whereas the latter shows that in the case of linear precoding with an average power constraint it is optimal to transmit in the directions of the channel eigenvectors. The case where $\boldsymbol{Q} = \boldsymbol{I}$ is of particular interest: This choice of mixing matrix $\boldsymbol{Q}$ combined with the optimal directivity matrix $\boldsymbol{U_F}$ corresponds to having orthogonal subchannels with no co-channel interference (cross-talk). This mode implies significantly reduced encoding and decoding complexity since each subchannel can be treated independently. Although orthogonal transmission is optimal in the sense of maximizing mutual information, it is not guaranteed to be optimal in the delay-limited case.

2.5.2 Non-linear precoding

For completeness, we note that an alternative to linear precoding is non-linear precoding, see [FWLH02b, HPS05] with references. Non-linear precoding techniques are commonly based on the Tomlinson–Harashima (TH) precoding strategy [HM72, Tom71]. The precoder inverts the channel, and uses modulus operations to reduce the transmitted power. Non-linear precoding is especially suitable for multi-user communication due to the ability to efficiently pre-eliminate intersymbol interference on the transmitter side, without the need for joint detection.

2.6 Discrete signal constellations

Because the complex-Gaussian signal constellation has a continuous distribution, its use is not an option for delay-limited transmission. In order to make the system practically implementable, we need to use discrete (and finite) signal-constellation sets. A discrete constellation is a finite set of points (typically in $\mathbb{R}$ or $\mathbb{C}$), where each point corresponds to a specific message or bit sequence[2]. Since the constellation points are separated in signal space, it is possible to estimate the exact sent message with high probability even though the signal is corrupted with noise.

Figure 2.3 shows common signal constellations in $\mathbb{C}$ of various types, representing bit rates from one up to six bits. One important factor that determines the probability of detection error is the minimum distance between constellation points, a dense signal constellation is more sensitive to noise than a sparse constellation because of the lower minimum distance. The binary phase-shift keying (BPSK) and the square quadrature amplitude modulated (QAM) constellations have an important property; they are linear in the real and imaginary parts. We can therefore construct a QAM constellation by linearly combining two real-valued pulse amplitude modulated (PAM) constellations.

[2]Finite codebooks with randomly generated Gaussian codeword blocks are also discrete, but with higher dimensions due to the concatenation of multiple symbols into codeword blocks.

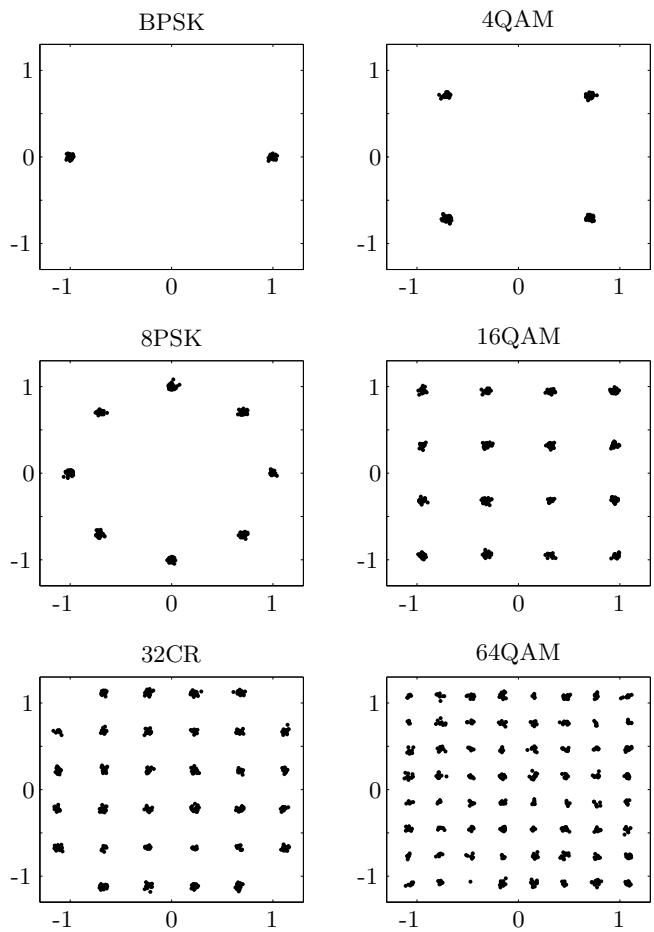

Figure 2.3: Signal constellations of various types. Note that the constellations 8PSK and 32CR are not linearly separable.

For the AWGN channel, the probability of detection error when using a QAM constellation can be tightly approximated as

$$P_e(\mathrm{SNR}) \simeq 4Q\left(\sqrt{\frac{3\,\mathrm{SNR}}{M-1}}\right), \tag{2.42}$$

where SNR denotes the signal to noise ratio and M denotes the number of points in the constellation. The function $Q(\cdot)$ is the Gaussian-tail function defined as

$$Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-y^2/2} dy. \tag{2.43}$$

When using linear precoding, different elements in $\boldsymbol{s}$ may be drawn from different signal constellations. If the channel has been orthogonalized, equation (2.42) can be used to determine the probability of detection error per symbol element. In the case when there is interference between the subchannels, the problem of computing the error probability depends on the type of detection algorithm that is used at the receiver.

2.6.1 The gap approximation

Equation (2.20) reveals that for Gaussian-distributed symbols there is a direct connection between data throughput and transmit power. The higher the power, the higher is the mutual information and the corresponding data rate. For discrete signal constellations where the data rate is determined by the number of points in the constellation, changing the power will mainly affect the error probability but in general not significantly alter the mutual information (or the information throughput). Hence, the number of constellation points needs to be optimized alongside the power optimization [FE91, GC97]. Deciding the optimal signal constellations, i.e. bit loading, is one of the main topics in this thesis.

Perhaps the most common approach for bit loading is to use the so-called gap approximation [CDEF95, PB05]: Use the orthogonalizing linear precoder, then use the constellations on the orthogonal subchannels according to the following bit rate (in nats)

$$b_i = \log\left(1 + \frac{\lambda_i p_i}{\Gamma}\right), \tag{2.44}$$

where $\Gamma \geq 1$ is denoted the SNR gap. The idea is that we can use the bit rate given by the capacity-optimal solution, but with a fixed penalty in terms of SNR given by the SNR gap Γ. By increasing the gap, the probability of a detection error reduces. From (2.42) we get the following relation between P_e and Γ

$$\Gamma = \frac{1}{3}\Big(Q^{-1}\big(P_e/4\big)\Big)^2. \tag{2.45}$$

The bit rate b_i cannot be an arbitrary real number, it must be rounded or discretized to match the set of available discrete signal constellations. However, it is

convenient (in terms of mathematical simplicity) to perform power optimization before rounding, rather than after. Maximizing the sum rate under a power constraint is very similar to the capacity-optimal power optimization problem. The optimal power is given by the water-filling solution

$$p_i = (\mu - \Gamma\lambda_i^{-1})^+, \tag{2.46}$$

where μ is determined such that the sum power satisfies $\sum_i p_i = P$. Insertion into (2.44) yields

$$b_i = \log\left(1 + \frac{\lambda_i p_i}{\Gamma}\right) = \left(\log\left(\frac{\lambda_i \mu}{\Gamma}\right)\right)^+ = \Big(\alpha + \log(\lambda_i)\Big)^+, \tag{2.47}$$

where α is another constant such that the sum rate equals the desired bit rate. These bit rates can now be discretized to match the set of available signal constellations as

$$b_i = \Big\lceil \alpha + \log(\lambda_i) \Big\rfloor^+, \qquad \sum_{i=1}^{N_t} b_i = R. \tag{2.48}$$

The bit loading defined by (2.48) is a good choice whenever we have orthogonal subchannels, this will be shown later in this thesis.

2.6.2 Gray coding

Up to now, we have not discussed how to map information, represented by a sequence of bits, to an element in the signal constellation set. Gray coding [Gra53] is a very efficient bit-to-constellation mapping for square QAM constellations. Figure 2.4 shows a Gray coding for 16-QAM. The main advantage with the Gray coding is that any point in the constellation differ by only one bit from its nearest neighbors. This means that when an detection error occurs, the number of bits that is detected incorrectly is in most cases only one or two bits. The bit error rate (BER) of a Gray coded symbol relates to the symbol error rate (SER) as

$$\text{BER} \approx \frac{\text{SER}}{b}, \tag{2.49}$$

where b denotes the number of bits represented by the constellation. For moderately low BERs, it can be shown that the dependency on b in the BER expression is dominated by the SER factor. Hence symmetric SER can serve as a good approximation to attain symmetric BERs as well.

2.7 Receiver structures

As was mentioned in Section 2.4, the delay-limited communication problem is not well defined without a specified detection algorithm. In this section we give an overview of the detection algorithms that will be considered in this work.

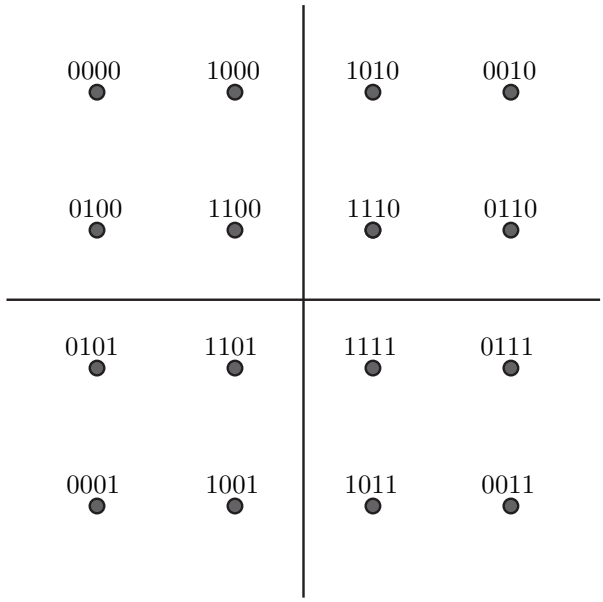

Figure 2.4: Gray coding for a 16-QAM signal constellation.

The task for the receiver is to detect the transmitted symbol vector $\boldsymbol{x}$ given the received signal $\boldsymbol{y}$ and the channel matrix $\boldsymbol{H}$. In systems employing iterative decoding between outer and inner codes, the a-priori information about $\boldsymbol{x}$ also serves as an input to the detector. The optimal detector is the maximum a-posteriori (MAP) detector, that maximizes the probability of $\boldsymbol{x}$ being sent given $\boldsymbol{y}$, $\boldsymbol{H}$, and the a-priori distribution of $\boldsymbol{x}$.

When outer codes are not taken into account, the a-priori distribution is uniform, and the MAP detector becomes the maximum likelihood (ML) detector that finds the vector $\boldsymbol{x}$ with the highest likelihood of observing $\boldsymbol{y}$, given that $\boldsymbol{x}$ was sent.

2.7.1 Maximum likelihood detector

When the noise vector has IID complex-Gaussian elements, the ML estimate is given by

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathcal{X}} |\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}|^2, \tag{2.50}$$

where $\mathcal{X}$ denotes the set of codewords. In general, this problem can be very difficult, cf. [JO05]. Herein, perfect CSI is available at the transmitter so the problem can be simplified by wisely designing the codebook, $\mathcal{X}$, based on the CSI.

2.7.2 Linear detector

The linear detector consists of a linear transformation of the received signal, subsequently followed by an element-wise (closest-point) detection of the signal constellations

$$\hat{\boldsymbol{x}} = \lceil \boldsymbol{W}^* \boldsymbol{y} \rfloor, \tag{2.51}$$

where $\lceil \cdot \rfloor$ denotes the element-wise closest-point detection. Strictly speaking this detector is non-linear since $\lceil \cdot \rfloor$ is non-linear, however, the joint multi-channel processing is a linear operation while the remaining (non-linear) detection is done element by element and not jointly, hence the name. Important special cases of this detector are the zero-forcing (ZF) detector

$$\boldsymbol{W}^* = (\boldsymbol{F}^* \boldsymbol{H}^* \boldsymbol{H} \boldsymbol{F})^{-1} \boldsymbol{F}^* \boldsymbol{H}^*, \tag{2.52}$$

and the minimum mean squared error (MMSE) detector

$$\boldsymbol{W}^* = \left(\boldsymbol{F}^* \boldsymbol{H}^* \boldsymbol{H} \boldsymbol{F} + \boldsymbol{I}\right)^{-1} \boldsymbol{F}^* \boldsymbol{H}^*. \tag{2.53}$$

The ZF detector removes all inter-symbol interference between subchannels but suffers from instability if $\boldsymbol{F}^* \boldsymbol{H}^* \boldsymbol{H} \boldsymbol{F}$ is badly conditioned. The MMSE detector is more robust at the price of a small bias in the estimate.

2.7.3 Decision feedback detector

The decision feedback (DF) detector uses a linear equalizer to detect the symbols of the subchannels one by one. Once a symbol is detected, its interference on the remaining subchannels (that are not yet detected) is removed. Figure 2.5 shows a schematic view of the DF detector. The received signal, $\boldsymbol{y}$, is passed to a linear filter, $\boldsymbol{w}_1$, to obtain an estimate, $\hat{s}_1$, of the first element of the transmitted signal vector $\boldsymbol{s}$. The estimate is passed to a closest-point detector to obtain the detected symbol $\tilde{s}_1$. Ideally the detected symbol is identical to the transmitted symbol of subchannel one. This enables us to remove all the inter-symbol interference that s_1 causes on the remaining subchannels. The linear filter, $\boldsymbol{w}_2^*$, gives an estimate of the second element of $\boldsymbol{s}$, subsequently all interference due to symbol s_1 is filtered out using the (scalar) feedback filter b_2^*. The resulting estimate, $\hat{s}_2$, is then passed on to a closest-point detector to obtain the detected symbol $\tilde{s}_2$ of the second subchannel. The procedure is repeated for all subchannels such that each detected subchannel is fed back to cancel out its interference on the remaining subchannels. The detection order has an impact on the performance, a rule of thumb is to detect the *strong* subchannels first to minimize the interference on the weaker subchannels. Herein however, because we perform joint bit loading and linear precoding, the detection order is automatically taken into account in the optimization.

Similar to the linear detector, the DF detector has two versions, the ZF or the MMSE detector. Linear detection is a special case of DF detection (use $\boldsymbol{B} = \boldsymbol{0}$).

Hence, if designed properly, the DF detector must be superior (or at least equivalent) to the linear detector. Note also that, even though the DF detector is slightly more complex than the linear detector, its complexity does not grow exponentially with the number of subchannels as is generally the case for the ML detection [JO05]. So, both in terms of complexity and performance we can regard the DF detector as an intermediate detector compared to the linear and ML detectors.

Instead of using the rather complicated subchannel-by-subchannel notation, we will use a vector–matrix notation illustrated to the right in Figure 2.5. To maintain causality (i.e. that no symbol is fed back that has not yet been detected) we enforce the condition that the feedback matrix $\boldsymbol{B}$ is strictly lower triangular.

2.8 Conclusion

In this chapter we introduced the MIMO system model, recapitulated the capacity-optimal transmission scheme, and formulated the problem of delay-limited transmission. We then introduced the linear precoding transmission strategy in combination with bit loading. Finally, Section 2.7 provided a short introduction to the three types of detection algorithms that will be considered in the thesis. We now continue to the two main parts of the thesis, starting with the linear precoding and bit loading design assuming ML detection.

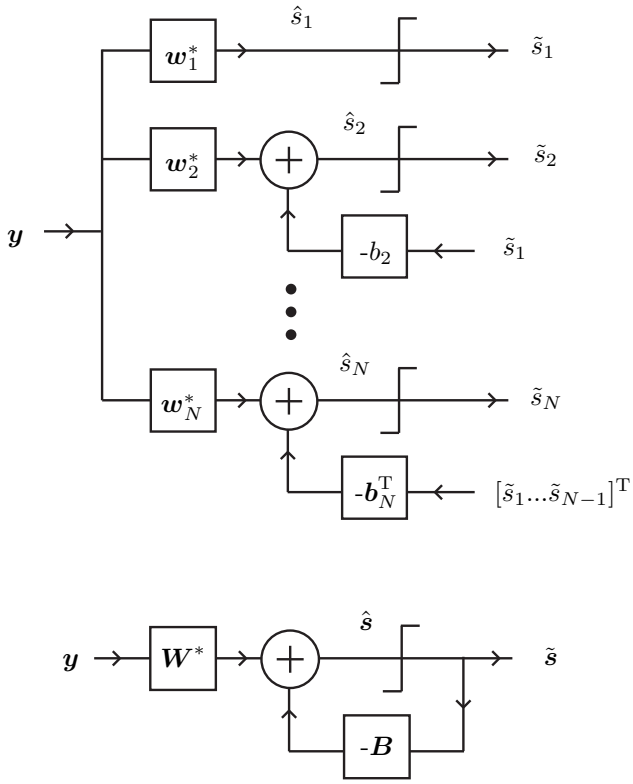

Figure 2.5: Schematic view of the DF detector.

Part I

Design based on maximum likelihood detection

Chapter 3

Introduction to Part I

In this part of the thesis we assume that the optimal maximum-likelihood (ML) detector [DGC03] is employed at the receiver. If data is blindly transmitted without any active effort to simplify the detection procedure, the computational complexity of ML detection grows exponentially with the number of dimensions in the MIMO system [JO05]. In fact, high complexity is the main motivation for using suboptimal detectors, such as the decision feedback detector that will be considered later in the thesis. However, if the transmitter has access to perfect CSI, we have the option to transmit the signals in such a way that the complexity of the optimal detection procedure can be significantly reduced. For instance, if we transmit data on orthogonal subchannels the complexity of ML detection is equal to that of the computationally inexpensive linear (zero-forcing) receiver.

Another use of TX-CSI is to improve system performance by turning off weak subchannels (to save power) and then send more data on the stronger subchannels. From information theory it is known that using TX-CSI improves the ergodic capacity of the system [Tel95], and it is reasonable to assume that there are potential gains for delay-limited transmission as well. We propose a transmission scheme using linear precoding with TX-CSI, that is designed explicitly for the ML detector with the main objective to be as power efficient as possible. Interestingly it turns out that it is suboptimal to orthogonalize the channel. It appears that the transmitter has to trade off decoding complexity at the receiver with performance in terms of power efficiency.

Essentially, ML detection involves an exhaustive search over all possible combinations of constellation points and then detects the most likely combination[1]. Because all combinations are tested, an ML detector can decode constellation points that are packed very densely in a high-dimensional signal space, which in turn allows for a reduced transmit power with an approximately unaltered block error rate. As a comparison, the ZF or MMSE detectors use a linear transformation to separate the subchannels, but the detection is only done subchannel by subchannel

[1] In practice, the search space can be reduced by using the so-called sphere decoder [DGC03].

(not jointly).

The design of signal constellations for the AWGN channel was considered in [For99] by Forney et al. Forney showed that the gain by using a constellation based on a lattice can be separated into a packing gain of the lattice, and a shaping gain of the constellation. Packing gain (also termed coding gain) is attained by packing points as densely as possible with fixed minimum distance, commonly referred to as sphere packing [CS88]. Shaping gain is determined by the shape of the union of all constellation points. Unfortunately, for high-dimensional systems it can be fairly demanding for the encoder as well as the decoder to use constellations with maximum shaping gain. Figure 3.1 illustrates the principles of coding and shaping gain for the AWGN channel. The leftmost constellation is a standard 64-QAM constellation. By making the constellation more circular while maintaining a fixed minimum distance between points, the average transmitted power can be reduced at approximately constant error probability. The gain due to the circular shape (as opposed to the square shape) is denoted the shaping gain of the constellation. In the figure the ideal shape is shown as a circle. The points in the QAM constellation is a subset of a square lattice. This square lattice is however not the densest packing in two dimensions; it is better to use the hexagonal lattice as illustrated in the rightmost constellation. By using a lattice with a denser sphere packing, less transmit power is required and we attain a so-called coding gain.

Consider the system equation of the parallel-SISO equivalent MIMO system (2.14). We see that the main difference between the MIMO channel and an AWGN channel (used multiple times), is the fact that channel gains differ between subchannels for the MIMO system. In our example from Figure 3.1 it could, for instance, be less costly to transmit in the vertical dimension compared to the horizontal dimension. Figure 3.2 shows a hypothetical case where a two-dimensional MIMO system has more gain in the vertical dimension. This asymmetry manifests into an elliptical ideal shaping region. Temporal variations imply that the optimal shape changes for every new channel realization. In the MIMO case, the problem of optimal constellation shaping is (even) more complex because the ideal shapes are elliptical (not circular). The 64-QAM constellation has a shape that is much worse for this MIMO channel compared to what it was for the AWGN channel in Figure 3.1, and we see that the importance of shaping is evident when there are asymmetric gains on the subchannels. The rightmost constellation illustrates that, similar to the AWGN case, by using a hexagonal lattice we attain a coding gain.

One drawback with constellations that are optimal in terms of shaping is that they are not linear, i.e., we can not generate the constellation as a linear combination of independent PAM symbols. Because of this non-linearity, the problem of mapping data to optimally shaped constellations is rather difficult if the number of dimensions is large. Since MIMO channels change over time, the ideal shaping regions change too. As a result the constellation mappings can not be precompiled, making it increasingly difficult to implement the system. The question we look into in this part of the thesis is: Can we obtain a linear code (easy to implement) that exploits some of the promised shaping and coding gain? Figure 3.3 demonstrates

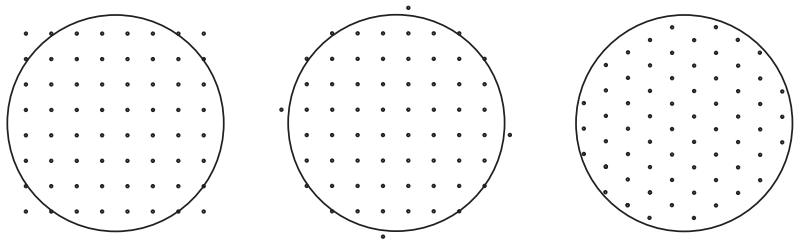

Figure 3.1: Shaping gain and coding gain for an AWGN channel.

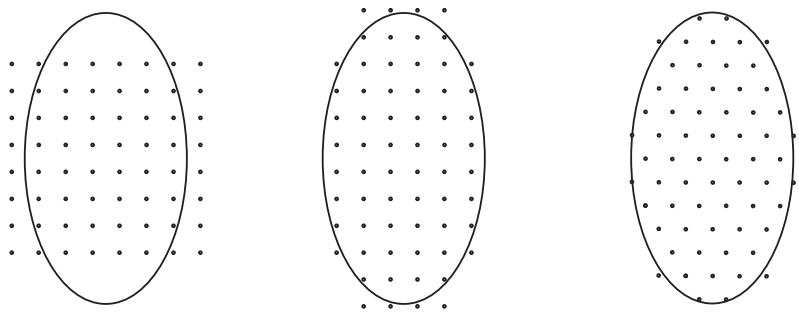

Figure 3.2: Shaping gain and coding gain for a MIMO channel.

three linear constellations for an asymmetric MIMO channel. The leftmost is the square 64-QAM constellation that we have seen before. The middle constellation is a rectangular QAM constellation that transmits 4 bits in the strong dimension and 2 bits in the weak dimension. This non-quadratic constellation represents a case of bit loading, i.e., the bit rate is adjusted to match the quality of the corresponding subchannel. However, in this example we can do even better by using a linear construction of the hexagonal lattice, as the rightmost constellation shows. Note that the rightmost constellation does not represent orthogonal transmission — the basis vectors in the generator matrix of the constellation are not perpendicular. Since the transmission is not orthogonal, linear detectors are suboptimal for this constellation.

In the following chapters, a close to optimal precoding scheme is presented that relies on known results concerning lattice sphere packing [CS88]. We propose to restrict the constellations to consist of linear combinations of PAM signals, and then optimize the shape of these constellations as far as possible. By using certain transformations that affect the constellation shape but not the underlying lattice structure, we can change the transmitted power at approximately constant block error rate (BLER). A two-step strategy to design close to optimal block codes under this restriction is proposed. The first step is to find the power optimal precoder for a fixed lattice structure. An approximate algorithm for this optimization is presented

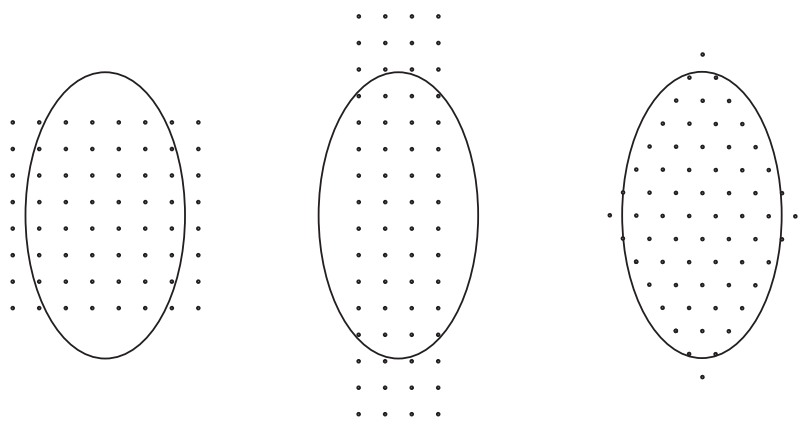

Figure 3.3: Linear constellations for a MIMO channel.

that combines lattice-aided basis reduction — similar to linear pre-equalization as proposed in [WF03] — with SVD precoding and adaptive bit loading. Lower and upper bounds on the minimized transmitted power that specify the limits of the algorithm are derived. The second step in the strategy is to determine which lattice structure that should be used as input to the algorithm. Essentially this corresponds to minimizing the lower bound, which turns out to be a problem related to lattice sphere packing. We will see that gains of several decibel (dB) can be achieved compared to the optimized diagonalizing precoder. A discussion regarding the suboptimality of the algorithm, due to approximations, is presented. In some sense, the presented precoding algorithm combines AWGN lattice coding [For99], traditional precoding with bit and power loading [BMO04, PB05], and lattice-aided basis reduction precoding [WF03] into one transmission scheme.

Chapter 4

The error probability and lattices

Exact performance analysis of the ML detector is a notoriously difficult problem. In many cases one has to resort to time-consuming Monte Carlo simulations of the entire MIMO system. This makes the task of designing the bit loading and precoder such that the performance is optimized an even more challenging problem. As of today, the optimal bit loading and precoder design using ML detection at the receiver is unknown. In this chapter we formulate a performance measure that approximates the probability of detection error and possesses sufficient mathematical simplicity for us to use it as optimization objective. In Chapter 5, we use this measure to develop a suboptimal joint bit loading and precoding scheme that improves the performance compared to the optimal orthogonal design by several dB.

4.1 System model

Consider the linear model of a discrete-time MIMO communication system, where N_t real-valued symbols are transmitted over a linear channel, from which N_r real-valued signals are received. In this part of the thesis, when dealing with ML detectors, all signals are real-valued since this simplifies the later notation significantly. Communication systems with complex signals can however always be reformulated on real form, as was shown in Section 2.1. The received signal block, $\boldsymbol{y} \in \mathbb{R}^{N_r}$, depends on the transmitted block, $\boldsymbol{x} \in \mathbb{R}^{N_t}$, as

$$\boldsymbol{y} = \boldsymbol{H}\,\boldsymbol{x} + \boldsymbol{n}, \tag{4.1}$$

where $\boldsymbol{H} \in \mathbb{R}^{N_r \times N_t}$ is the channel matrix and $\boldsymbol{n} \in \mathbb{R}^{N_r}$ is an additive Gaussian noise vector. It is assumed that the noise distribution is known, and that the noise component has already been linearly pre-whitened. Furthermore, without loss of generality, the signal is normalized such that the noise variance is unity, and $\boldsymbol{n}$ is consequently distributed as $\boldsymbol{n} \sim N(\boldsymbol{0}, \boldsymbol{I}_{N_r})$.

Denote Q as the rank of the channel matrix, then define the singular value decomposition (SVD) of the channel matrix as $\boldsymbol{H} = \boldsymbol{U}_{\boldsymbol{H}}\boldsymbol{\Lambda}_{\boldsymbol{H}}\boldsymbol{V}_{\boldsymbol{H}}^{\mathrm{T}}$, where both

$\boldsymbol{U_H} \in \mathbb{R}^{N_r \times Q}$ and $\boldsymbol{V_H} \in \mathbb{R}^{N_t \times Q}$ have orthonormal columns, and $\boldsymbol{\Lambda_H} \in \mathbb{R}^{Q \times Q}$ is diagonal and nonsingular. No information about $\boldsymbol{x}$ in $\boldsymbol{y}$ is lost when projecting the received signal to Q dimensions as

$$\tilde{\boldsymbol{y}} = \boldsymbol{U_H}^{\mathrm{T}} \boldsymbol{y} = \boldsymbol{\Lambda_H} \boldsymbol{V_H}^{\mathrm{T}} \boldsymbol{x} + \tilde{\boldsymbol{n}}, \tag{4.2}$$

with the noise $\tilde{\boldsymbol{n}} \sim N(\boldsymbol{0}, \boldsymbol{I}_Q)$. A PAM linear dispersion code can be formulated as $\boldsymbol{V_H}^{\mathrm{T}} \boldsymbol{x} = \boldsymbol{F}\, \boldsymbol{s}$, where $\boldsymbol{F} \in \mathbb{R}^{Q \times N}$ is a data-independent precoding matrix, and $\boldsymbol{s}$ is the PAM data-symbols vector. It is assumed that the number of subchannels, N, has been chosen such that $N \leq Q$. The elements of $\boldsymbol{s}$ are assumed to be independent, zero-mean random variables representing the data to be conveyed over the channel.

4.2 The union bound

Our objective is to optimize the system performance. The first step is to derive an expression of the probability of detection error. Later we will use this expression to search for the Pareto optimum between data rate, transmitted power and detection error probability. Based on the system model (4.1) we can formulate the ML detection criterion (2.50) as

$$\hat{\boldsymbol{s}} = \arg \min_{\boldsymbol{s} \in \mathbb{S}} |\tilde{\boldsymbol{y}} - \boldsymbol{\Lambda_H} \boldsymbol{F} \boldsymbol{s}|^2, \tag{4.3}$$

where $\mathbb{S}$ denotes the set of possible symbol vectors. The probability of detection error, P_e, is equal to one minus the probability of correct detection

$$P_e = 1 - \Pr\left\{ \bigcup_{\substack{\hat{\boldsymbol{s}}, \boldsymbol{s} \in \mathbb{S} \\ \hat{\boldsymbol{s}} \neq \boldsymbol{s}}} |\boldsymbol{\Lambda_H} \boldsymbol{F}(\boldsymbol{s} - \hat{\boldsymbol{s}}) + \tilde{\boldsymbol{n}}|^2 > |\tilde{\boldsymbol{n}}|^2 \right\}, \tag{4.4}$$

which unfortunately is difficult to put in closed form. Instead, we use the well known union bound [LS03] to get an upper bound[1]

$$P_e \leq \sum_{\boldsymbol{s} \in \mathbb{S}} \frac{1}{2^R} \sum_{\substack{\hat{\boldsymbol{s}} \in \mathbb{S} \\ \hat{\boldsymbol{s}} \neq \boldsymbol{s}}} \Pr\left\{ |\boldsymbol{\Lambda_H} \boldsymbol{F}(\boldsymbol{s} - \hat{\boldsymbol{s}}) + \boldsymbol{n}|^2 < |\boldsymbol{n}|^2 \right\}, \tag{4.5}$$

that consists of a sum of pairwise error probabilities (PEP) defined as

$$\mathrm{PEP}_{\hat{\boldsymbol{s}}, \boldsymbol{s}} = \Pr\left\{ |\boldsymbol{\Lambda_H} \boldsymbol{F}(\boldsymbol{s} - \hat{\boldsymbol{s}}) + \boldsymbol{n}|^2 < |\boldsymbol{n}|^2 \right\}. \tag{4.6}$$

By observing that the difference between the two norms in (4.6) is normal distributed

$$|\boldsymbol{\Lambda_H} \boldsymbol{F}(\boldsymbol{s} - \hat{\boldsymbol{s}}) + \boldsymbol{n}|^2 - |\boldsymbol{n}|^2 \sim N\Big(|\boldsymbol{\Lambda_H} \boldsymbol{F}(\boldsymbol{s} - \hat{\boldsymbol{s}})|^2, 4|\boldsymbol{\Lambda_H} \boldsymbol{F}(\boldsymbol{s} - \hat{\boldsymbol{s}})|^2 \Big), \tag{4.7}$$

[1] Note that since all symbols are equally likely, the a-priori probability that a specific codeword was transmitted is $|\mathbb{S}|^{-1} = 2^{-R}$, where R is the rate.

we can compute the PEPs as

$$\mathrm{PEP}_{\hat{\boldsymbol{s}},\boldsymbol{s}} = Q\left(\frac{|\boldsymbol{\Lambda_H F}(\hat{\boldsymbol{s}}-\boldsymbol{s})|}{2}\right), \tag{4.8}$$

where the Q-function is the Gaussian-tail integral given by (2.43). The PEP formula shows that the union bound is determined by the set of all difference vectors $\boldsymbol{\Lambda_H F}(\hat{\boldsymbol{s}}-\boldsymbol{s})$. In the next section we show how this set is a subset of a lattice, which will allow us to formulate a very compact approximation of the error probability.

4.3 The error probability as a function of a lattice

Although the union bound (4.5) is an approximation of the true error probability, the bound is not easy to analyze in closed form. In order to obtain a cost function that is easier to optimize we will simplify (4.5) by making appropriate approximations. The first step is to show that the difference vectors constitute a subset of a lattice. The points in a b-bit PAM constellation can be defined as

$$s = \sqrt{\frac{12}{4^b-1}}\left(z - \frac{2^b-1}{2}\right), \tag{4.9}$$

where z is an integer in the range $0, ..., 2^b - 1$. Note that the normalization is such that the constellation uses unit power on average. The difference between two points in a b-bit PAM constellation is therefore a subset of

$$\hat{s} - s \in \left\{ \sqrt{\frac{12}{4^b-1}}\, z \;\middle|\; z \in \mathbb{Z} \right\}, \tag{4.10}$$

which is, in fact, a one-dimensional lattice. Extending this to N dimensions yields the set of difference vectors

$$\hat{\boldsymbol{s}} - \boldsymbol{s} \in \left\{ \boldsymbol{\Sigma}^{-1}\boldsymbol{z},\; \middle|\; \boldsymbol{z} \in \mathbb{Z}^N \right\}, \tag{4.11}$$

where the diagonal normalization matrix, $\boldsymbol{\Sigma}$, is defined as

$$\boldsymbol{\Sigma} = \mathrm{diag}\left\{ \sqrt{\frac{4^{b_1}-1}{12}}, ..., \sqrt{\frac{4^{b_N}-1}{12}} \right\}, \tag{4.12}$$

and where $b_1, ..., b_N$ denotes the bit load on each element.

By parameterizing the difference vectors by the corresponding integer vectors, $\boldsymbol{z} \in \mathbb{Z}^N$, we can reformulate the union bound as

$$\bar{P}_e = \sum_{\boldsymbol{z}\in\mathbb{Z}^N,\, \boldsymbol{z}\neq\boldsymbol{0}} \frac{A(\boldsymbol{z})}{2^R} Q\left(\frac{|\boldsymbol{\Lambda_H F \Sigma}^{-1}\boldsymbol{z}|}{2}\right), \tag{4.13}$$

where $A(\boldsymbol{z})$ denotes the number of terms in (4.5) that has a difference vector corresponding to $\boldsymbol{z}$. By counting the number of terms for a specific $\boldsymbol{z}$, we get

$$A(\boldsymbol{z}) = \begin{cases} \prod_{k=1}^{N} \left(2^{b_k} - |z_k|\right) & \text{if } |z_k| \leq 2^{b_k} \ \forall\ k = 1, ..., N \\ 0 & \text{otherwise} \end{cases}. \tag{4.14}$$

By inspection, the number of terms can be upper bounded as $A(\boldsymbol{z}) < 2^R$, and thus, (4.13) can be upper bounded as

$$\bar{P}_e < \sum_{\boldsymbol{z} \in \mathbb{Z}^N,\, \boldsymbol{z} \neq \boldsymbol{0}} Q\left(\frac{|\boldsymbol{\Lambda_H F \Sigma}^{-1}\boldsymbol{z}|}{2}\right). \tag{4.15}$$

The main problem with applying the upper bound $A(\boldsymbol{z}) < 2^R$ is that if $\boldsymbol{\Lambda_H F \Sigma}^{-1}$ is not well conditioned, then there may exist vectors $\boldsymbol{z}$ such that $A(\boldsymbol{z}) = 0$ but for which $Q(|\boldsymbol{\Lambda_H F \Sigma}^{-1}\boldsymbol{z}|/2)$ is dominating in (4.15). Here we postulate that (4.15) is approximately tight in the high SNR region when $\boldsymbol{\Lambda_H F \Sigma}^{-1}$ is sufficiently well conditioned.

It is clear that we can express the vectors $\boldsymbol{\Lambda_H F \Sigma}^{-1}\boldsymbol{z}$ in (4.15) as points in a lattice with generator matrix $\boldsymbol{M} = \boldsymbol{\Lambda_H F \Sigma}^{-1}$. Due to the steep descent of the Q-function towards zero, only a few terms in the union bound contribute to $\bar{P}_e$, namely the terms containing small Euclidian norms $|\boldsymbol{Mz}|$. The smallest norm,

$$\xi(\boldsymbol{M}) = \min_{\boldsymbol{z} \in \mathbb{Z}^N,\, \boldsymbol{z} \neq \boldsymbol{0}} \sqrt{\boldsymbol{z}^{\mathrm{T}} \boldsymbol{M}^{\mathrm{T}} \boldsymbol{M} \boldsymbol{z}}, \tag{4.16}$$

is referred to as the minimum distance of the lattice. Typically there are multiple vectors with norms of length $\xi(\boldsymbol{M})$. We therefore define the kissing number, $K(\boldsymbol{M})$, of the lattice as the number of distinct vectors $\boldsymbol{z} \in \mathbb{Z}^N$ such that $|\boldsymbol{Mz}| = \xi(\boldsymbol{M})$. (See [CS88] for more on kissing numbers.) By assuming that only the $K(\boldsymbol{M})$ minimum-distance terms contribute to (4.15), the probability of a detection error can be approximated as

$$P_e \approx K(\boldsymbol{M}) Q\left(\frac{\xi(\boldsymbol{M})}{2}\right). \tag{4.17}$$

Since the lattice is a linear construction, the kissing number is also the number of nearest neighbors to any point in the lattice. To illustrate this, the minimum distance and kissing number of the hexagonal E_2 lattice is shown in Figure 4.1. Note that (4.17) relies on high SNR assumptions, therefore the approximation is not valid for low SNR per bits or when $\boldsymbol{M}$ is ill-conditioned.

Given that the CSI at the transmitter is perfect, the problem of designing a linear precoder for ML detection is equivalent to the problem of choosing a lattice generator matrix $\boldsymbol{M}$. The choice of lattice affects the probability of error via the minimum distance and the kissing number, but we also have constraints on $\boldsymbol{M}$ through the limitations on the transmit power and the data rate. As an example, consider using the $\mathbb{Z}^N$ lattice, where the generator matrix is a scaled identity matrix

$\boldsymbol{M} = \boldsymbol{I}\xi$. This lattice corresponds to an orthogonalizing precoder since the effective channel is interference free. The lattice has a minimum distance $\xi(\boldsymbol{M}) = \xi$ and a kissing number $K(\boldsymbol{M}) = 2N$. The transmit power is given by $P = \xi^2 \mathrm{Tr}\{\boldsymbol{\Lambda}_{\boldsymbol{H}}^{-2}\boldsymbol{\Sigma}^2\}$. In order to maximize the performance we need to distribute the bit loading so that the total data rate R remains fixed while

$$\frac{P}{\xi^2} = \mathrm{Tr}\{\boldsymbol{\Lambda}_{\boldsymbol{H}}^{-2}\boldsymbol{\Sigma}^2\} \tag{4.18}$$

is minimized. This allows us to either minimize the probability of error (by keeping P fixed), or minimize the transmitted power P (by keeping ξ^2 fixed). In the next chapter we will generalize this precoder synthesis strategy to arbitrary lattices.

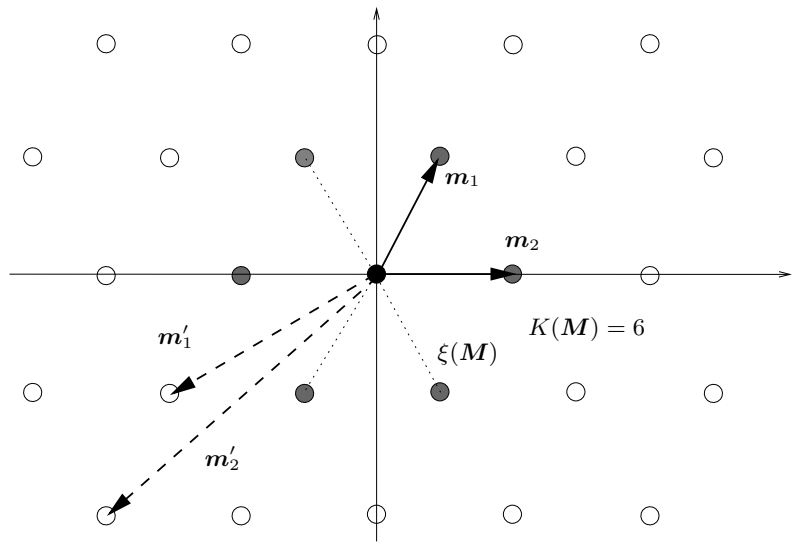

Figure 4.1: An illustration of the hexagonal E_2 lattice. The generator matrix $\boldsymbol{M}$ with column vectors $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ fully defines all lattice points. So do the basis vectors $\boldsymbol{m}_1'$ and $\boldsymbol{m}_2'$ that can be calculated by multiplying $\boldsymbol{M}$ with an integer matrix $\boldsymbol{B}$. The kissing number (which in this case is equal to six) and the minimum distance between any points in the lattice are denoted $K(\boldsymbol{M})$ and $\xi(\boldsymbol{M})$ respectively. These properties are not only invariant to basis vector changes, but also to rotations of the entire $\mathbb{R}^2$ space using an orthogonal matrix $\boldsymbol{U}$.

Chapter 5

Lattice-based precoding

An interesting property of the performance measure (4.17) is that it is completely determined by the lattice represented by the generator matrix, $\boldsymbol{M}$. This allows us to use well-known results regarding lattices in the search for the optimal precoder. In particular we will use that lattices are invariant to rotations and changes of basis vectors. Let $\boldsymbol{U}$ be a real-valued, (full column-rank) matrix with orthonormal columns, and let $\boldsymbol{B}$ be a unimodular integer-valued matrix with determinant $|\boldsymbol{B}| = \pm 1$. When multiplying the generator matrix with $\boldsymbol{U}$ and $\boldsymbol{B}$, on the left and right side respectively, the new generator matrix represents the same lattice as the previous matrix. In particular, the minimum distance and the kissing number remain unchanged as $\xi(\boldsymbol{UMB}) = \xi(\boldsymbol{M})$, and $K(\boldsymbol{UMB}) = K(\boldsymbol{M})$. The matrix $\boldsymbol{U}$ rotates the lattice points around the origin, and will therefore be denoted the rotation matrix. The matrix $\boldsymbol{B}$ changes the basis vectors of the lattice generator matrix, $\boldsymbol{M}$, but not the lattice itself. The matrix will be referred to as the basis reduction matrix. Figure 4.1 demonstrates how two sets of basis vectors can represent the same lattice.

Assume that we have selected an initial lattice generator matrix, $\boldsymbol{M}_0 \in \mathbb{R}^{N\times N}$, that results in a desired probability of error. Using a rotation matrix and a basis reduction matrix we create a new generator matrix $\boldsymbol{M}$ that relates to the precoder as

$$\boldsymbol{F} = \boldsymbol{\Lambda}_{\boldsymbol{H}}^{-1}\boldsymbol{M}\boldsymbol{\Sigma} = \boldsymbol{\Lambda}_{\boldsymbol{H}}^{-1}\boldsymbol{U}\boldsymbol{M}_0\boldsymbol{B}\boldsymbol{\Sigma}, \tag{5.1}$$

where $\boldsymbol{U} \in \mathbb{R}^{Q\times N}$ has orthonormal columns and $\boldsymbol{B} \in \mathbb{Z}^{N\times N}$ is integer-valued and unimodular. Because of the lattice equivalence properties we are essentially free to modify $\boldsymbol{U}$ and $\boldsymbol{B}$, as well as the bit load $\boldsymbol{\Sigma}$ arbitrarily without significantly affecting the error probability. What will be affected is the transmit power

$$P = \mathrm{Tr}\{\boldsymbol{\Lambda}_{\boldsymbol{H}}^{-2}\boldsymbol{U}\boldsymbol{M}_0\boldsymbol{B}\boldsymbol{\Sigma}^2\boldsymbol{B}^{\mathrm{T}}\boldsymbol{M}_0^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\}. \tag{5.2}$$

This allows us to reduce the transmitted power at some fixed BLER, making the precoder more efficient. The BLER approximation (4.17) is monotonically decreasing with the transmitted power, P. Hence, minimizing the transmit power and

minimizing BLER is essentially the same problem — the difference is simply a re-scaling of the generator matrix. In other words, the optimum is a Pareto optimum.

5.1 Precoding algorithm

Given a nonsingular generator matrix, $\boldsymbol{M}_0 \in \mathbb{R}^{N\times N}$, our goal is to find the most power efficient precoder, $\boldsymbol{F}$. In order not to confuse $\boldsymbol{M}_0$ with $\boldsymbol{M}$, the matrix $\boldsymbol{M}_0$ will be denoted as the lattice base in this work. Minimizing the power (5.2) involves joint optimization of both discrete and continuous variables. This problem is difficult (if not impossible) to solve optimally. In the following subsections a sub-optimal algorithm is proposed: The idea is to optimize each parameter sequentially, and iterate until the solution converges.

5.1.1 Optimization of the rotation matrix

The optimization of the rotation matrix, $\boldsymbol{U}$, for fixed $\boldsymbol{M}_0$, $\boldsymbol{B}$, and $\boldsymbol{\Sigma}$ corresponds to solving an SVD. The procedure is summarized in the following theorem:

Theorem 5.1.1 *The power minimizing rotation matrix, $\boldsymbol{U}$, for a fixed $\boldsymbol{M}_0$, $\boldsymbol{B}$, and $\boldsymbol{\Sigma}$ is*

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{V}^{\mathrm{T}} \\ \boldsymbol{0}_{(Q-N)\times N} \end{bmatrix}, \tag{5.3}$$

where $\boldsymbol{V}$ is given by the SVD of

$$\boldsymbol{M}_0\boldsymbol{B}\boldsymbol{\Sigma}^2\boldsymbol{B}^{\mathrm{T}}\boldsymbol{M}_0^{\mathrm{T}} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}. \tag{5.4}$$

Note that the singular values $\boldsymbol{\Lambda}$ are (per definition) non-increasing along the diagonal.

Proof: Define

$$\begin{aligned}\boldsymbol{\lambda}_1 &= \mathrm{diag}\{\boldsymbol{\Lambda}_{\boldsymbol{H}}^{-2}\},\\ \boldsymbol{\lambda}_2 &= [\mathrm{diag}\{\boldsymbol{\Lambda}\}^{\mathrm{T}}\ \boldsymbol{0}_{1\times Q-N}]^{\mathrm{T}}.\end{aligned}$$

The transmitted power is then $P = \boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\lambda}_2$, where $\boldsymbol{Q} \in \mathbb{R}^{Q\times Q}$ is a doubly stochastic matrix that depends on $\boldsymbol{U}$. By Birkhoff's Theorem (see [HJ85]) we know that all doubly stochastic matrices can be decomposed into a finite linear combination of all permutation matrices, $\boldsymbol{\Pi}_1, ..., \boldsymbol{\Pi}_{Q!}$, of the same dimensions. More specifically,

$$\boldsymbol{Q} = \sum_{n=1}^{Q!} \boldsymbol{\Pi}_n \alpha_n^2,$$

where $\sum_n \alpha_n^2 = 1$. Minimizing $P = \sum_{n=1}^{Q!} \left(\boldsymbol{\lambda}_1^{\mathrm{T}} \boldsymbol{\Pi}_n \boldsymbol{\lambda}_2\right) \alpha_n^2$ subject to $\sum_n \alpha_n^2 = 1$ gives $\boldsymbol{Q} = \boldsymbol{\Pi}_m$ where

$$m = \arg \min_n \boldsymbol{\lambda}_1^{\mathrm{T}} \boldsymbol{\Pi}_n \boldsymbol{\lambda}_2.$$

The power is minimized if we let the i'th largest element of $\boldsymbol{\lambda}_1$ be multiplied with the i'th smallest element of $\boldsymbol{\lambda}_2$ for all i. Since $\boldsymbol{\lambda}_1$ is increasing with the indices and $\boldsymbol{\lambda}_2$ is decreasing with the indices, we conclude that the optimum is achieved with $\boldsymbol{Q} = \boldsymbol{I}_Q$, which is equivalent to choosing $\boldsymbol{U}$ as in (5.3). □

5.1.2 Optimization of the bit load

There are numerous algorithms available for multi-carrier bit loading in the literature [HH, CCB95, FH96, KRJ98, Cam99]. Most of them assume orthogonal subchannels and use the gap approximation [CDEF95] to maximize the mutual information of the system. In this work, the focus is on maximizing the minimum distance rather than the mutual information. Moreover, our problem differs because of the potential cross talk between the subchannels. The following algorithm optimally redistributes the bit load such that the power is minimized, while the total data rate, R, and the matrices $\boldsymbol{U}$, $\boldsymbol{M}_0$, and $\boldsymbol{B}$ are kept fixed.

Introduce the transmit-signal basis matrix as

$$\boldsymbol{G} = \boldsymbol{\Lambda}_{\boldsymbol{H}}^{-1} \boldsymbol{U} \boldsymbol{M}_0 \boldsymbol{B}, \tag{5.5}$$

and denote $g_1, ..., g_N$, such that g_i is the i'th diagonal element of $\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}$. The power consumption is then

$$P = \sum_i \frac{g_i(4^{b_i} - 1)}{12}. \tag{5.6}$$

If one bit is moved from subchannel n to m, the change in power consumption will be

$$P_{\mathrm{new}} - P_{\mathrm{old}} = -\frac{g_n 4^{b_n}}{16} + \frac{g_m 4^{b_m}}{4}. \tag{5.7}$$

Hence, in order to reduce the power consumption, n and m have to be selected such that $g_n 4^{b_n} > 4\, g_m 4^{b_m}$. Using this result, we propose the following algorithm for bit loading:

Theorem 5.1.2 *Algorithm 5.1.2 finds the global optimum with respect to $b_1, ..., b_N$.*

Proof: See Appendix 5.A. □

5.1.3 Optimization of the basis reduction matrix

The basis reduction matrix, $\boldsymbol{B}$, is more difficult to optimize compared to the two previous cases. There do however exist efficient but sub-optimal algorithms for

Initialize $b_1, ..., b_N$ by distributing the R bits as evenly as possible among the sub-channels.

1. Let n be the index of the maximum element of $g_1 4^{b_1}, ..., g_N 4^{b_N}$ with non-zero bit load. Then let m be the index of the minimum element.
2. Check whether $g_n 4^{b_n} > 4\, g_m 4^{b_m}$. If so, move one bit from b_n to b_m and go to 1. Otherwise, the bit load has been optimized and the algorithm can terminate.

lattice basis reduction that reduce the transmitted power. One such algorithm that has polynomial average complexity is the Lenstra, Lenstra and Lovasz (LLL) algorithm [LLL82, JSM08]. Here, the LLL algorithm is used to reduce the basis of the transmit signal basis matrix $\boldsymbol{G}_0 = \boldsymbol{\Lambda}_{\boldsymbol{H}}^{-1}\boldsymbol{U}\boldsymbol{M}_0$. The result is a new basis, $\boldsymbol{G} = \boldsymbol{G}_0\boldsymbol{B}$, that represents an equivalent lattice, but where the lengths of the basis vectors have been reduced. For this to hold, the matrix $\boldsymbol{B}$ has to be integer-valued with determinant $|\boldsymbol{B}| \pm 1$.

The main idea of the LLL algorithm is to make the basis vectors as orthogonal as possible. A good measure of orthogonality is to compare the product of the lengths of the basis vectors with the determinant as

$$D(\boldsymbol{G}_0) = \frac{\prod_{i=1}^{N}[\boldsymbol{G}_0^{\mathrm{T}}\boldsymbol{G}_0]_{i,i}^{1/N}}{|\boldsymbol{G}_0^{\mathrm{T}}\boldsymbol{G}_0|^{1/N}}. \tag{5.8}$$

Due to the Hadamard inequality [Had93], $D(\boldsymbol{G}_0)$ is always greater than or equal to one. As $D(\boldsymbol{G}_0)$ approaches one, the basis $\boldsymbol{G}_0$ becomes more orthogonal. The following upper bound on the reduced basis was derived in [LLL82]

$$D(\boldsymbol{G}_0\boldsymbol{B}) \leq 2^{(N-1)/4}. \tag{5.9}$$

However, empirical experience suggests that the bound is typically not so tight: For most initial bases $\boldsymbol{G}_0$ of practical interest, $D(\boldsymbol{G}_0\boldsymbol{B})$ is closer to one than the upper bound. The orthogonality measure will play an important role later in this work when we discuss performance bounds on the optimization algorithm.

5.1.4 Combined optimization

When combining the above optimization procedures, one has to decide in what order and how often each procedure should be performed. After comparing various configurations experimentally, we believe that the ordering is not of major significance to the final result. In the later simulations we used the procedure described here: First, initialize the basis reduction matrix as $\boldsymbol{B} = \boldsymbol{I}$, distribute the bit load as evenly as possible, and perform an initial optimization of the rotation matrix, $\boldsymbol{U}$. The algorithm should then iteratively optimize the matrices in the following

order: $\boldsymbol{B}$, $\boldsymbol{U}$, $\boldsymbol{\Sigma}$, $\boldsymbol{U}$. The iteration should stop when the algorithm converges, and no further improvements are obtained. By optimizing $\boldsymbol{U}$ in between every $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ optimization step, it appears that convergence speed is improved. A reason for this could be that $\boldsymbol{U}$ is a continuous entity. Its optimization is more well behaved without the flip-flop characteristics of the discrete optimization steps, $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$.

Sub-channels that have zero bit load in the optimization should be switched off by removing the corresponding columns from the precoder matrix, $\boldsymbol{F}$. For later notational simplicity, we now redefine N as the number of subchannels with non-zero bit loads $b_1, ..., b_N$ after the combined power optimization has been performed. Furthermore, using (5.3), we define $\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H},N} \in \mathbb{R}^{N\times N}$ using the SVD of $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{H}}^2\boldsymbol{U}$ as

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{H}}^2\boldsymbol{U} = \boldsymbol{V}\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H},N}^2\boldsymbol{V}^{\mathrm{T}}, \tag{5.10}$$

where $\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H},N}$ is diagonal and contains the upper left sub-matrix of the $\boldsymbol{\Lambda}_{\boldsymbol{H}}$ matrix.

5.2 Bounds on the performance

Using the three invariance properties described in Section 5.1, it is possible to reduce the transmit power without drastically affecting the BLER. A valid question here is: To what extent can the power be reduced in this way? Clearly, because the error probability remains fixed, the transmitted power must at least be greater than zero. This section shows that, for a fixed lattice base, there exists a fundamental lower bound on the transmitted power. The lower bound is independent of the bit load distribution[1], as well as the matrices $\boldsymbol{U}$, and $\boldsymbol{B}$. It can therefore be used as a goal, or benchmark, for the power-minimizing algorithm presented above. Furthermore, the bound also provides insight into the problem on how to optimally select the lattice base matrix, $\boldsymbol{M}_0$.

5.2.1 Lower bound

Using the fact that an algebraic mean of positive entities is larger than or equal to the geometric mean, it can be established that the transmit power is bounded as

$$P = \mathrm{Tr}\{\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}\} \geq N\prod_{i=1}^{N}[\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}]_{i,i}^{1/N}. \tag{5.11}$$

The bound is further refined using Hadamard's inequality

$$\begin{aligned} P &\geq N\left|\boldsymbol{B}^{\mathrm{T}}\boldsymbol{M}_0^{\mathrm{T}}\boldsymbol{V}\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H},N}^{-2}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{M}_0\boldsymbol{B}\boldsymbol{\Sigma}^2\right|^{1/N} \\ &= N\left(\frac{|\boldsymbol{M}_0|^2|\boldsymbol{\Sigma}^2|}{|\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H},N}|^2}\right)^{1/N} \\ &\geq \frac{N4^{R/N}}{16}\cdot\frac{|\boldsymbol{M}_0|^{2/N}}{|\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H},N}|^{2/N}}, \end{aligned} \tag{5.12}$$

[1]The bound is however highly dependent on the number of active substreams.

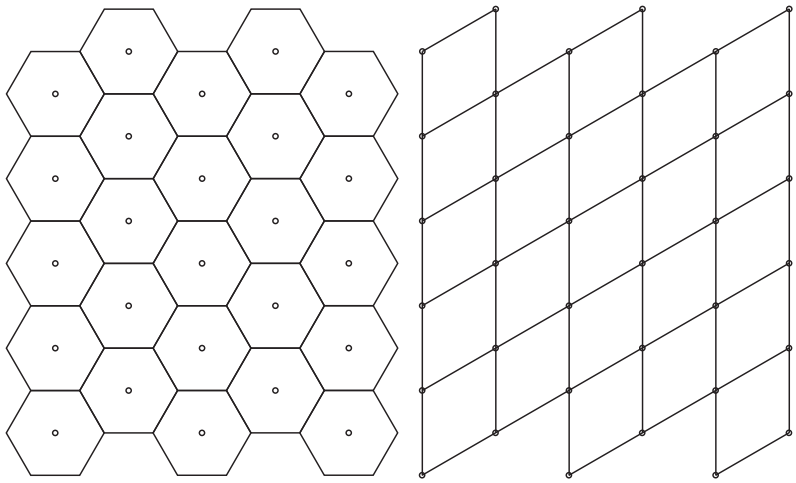

Figure 5.1: Relation between the density of the lattice and the determinant of the generator matrix. The left-hand side shows the Voronoi cells of the E_2 lattice. The right-hand side shows the parallelotopes spanned by the basis vectors of the same lattice. Clearly, the hyper-dimensional volume (in this case the area) of the Voronoi cell is equal to the volume of the parallelotopes, which is equal to the determinant of the generator matrix.

where the last inequality is due to

$$|\boldsymbol{\Sigma}^2|^{1/N} = \frac{\prod_{i=1}^{N} 4^{b_i/N}(1-4^{-b_i})^{1/N}}{12} \geq \frac{4^{R/N}}{16}, \tag{5.13}$$

since all bit loads $b_i \geq 1$. Note that, because the right hand side of (5.12) is independent of $\boldsymbol{B}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{U}$, the bound is a definite lower bound on the minimized transmit power.

In order to apply the bound to the BLER expression (4.17), we need to express it in terms of the minimum distance of the lattice. The determinant, $|\boldsymbol{M}_0|$, is the volume of a Voronoi cell of the lattice; it is related to the minimum distance of the lattice by the packing gain, $\sigma(\boldsymbol{M}_0)$, defined here as[2]

$$\xi^2(\boldsymbol{M}_0) = |\boldsymbol{M}_0|^{2/N}\sigma^2(\boldsymbol{M}_0). \tag{5.14}$$

Figure 5.1 shows the relation between the density of the lattice, the Voronoi cell, and the determinant of the Gram matrix, $\boldsymbol{M}_0^{\mathrm{T}}\boldsymbol{M}_0$. For a fixed minimum distance, the density of the lattice (i.e., the packing gain) is determined by the volume of the Voronoi cells: The smaller volume, the denser the lattice. The figure illustrates

[2]See also [CS88].

$$\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M} = \begin{bmatrix} 4 & 0 & 0 & -2 & 0 & 0 & 2 & -1 & -1 & -1 & 2 & -1 \\ 0 & 4 & 0 & 0 & -2 & 0 & 2 & -1 & -1 & -1 & -1 & 2 \\ 0 & 0 & 4 & 0 & 0 & -2 & 2 & 2 & 2 & -1 & -1 & -1 \\ -2 & 0 & 0 & 4 & 0 & 0 & -1 & -1 & 2 & 2 & -1 & -1 \\ 0 & -2 & 0 & 0 & 4 & 0 & -1 & 2 & -1 & 2 & -1 & -1 \\ 0 & 0 & -2 & 0 & 0 & 4 & -1 & -1 & -1 & 2 & 2 & 2 \\ 2 & 2 & 2 & -1 & -1 & -1 & 4 & 0 & 0 & -2 & 0 & 0 \\ -1 & -1 & 2 & -1 & 2 & -1 & 0 & 4 & 0 & 0 & -2 & 0 \\ -1 & -1 & 2 & 2 & -1 & -1 & 0 & 0 & 4 & 0 & 0 & -2 \\ -1 & -1 & -1 & 2 & 2 & 2 & -2 & 0 & 0 & 4 & 0 & 0 \\ 2 & -1 & -1 & -1 & -1 & 2 & 0 & -2 & 0 & 0 & 4 & 0 \\ -1 & 2 & -1 & -1 & -1 & 2 & 0 & 0 & -2 & 0 & 0 & 4 \end{bmatrix}.$$

Figure 5.2: Example of a 12 dimensional lattice. The equation shows a Gram matrix for the K_{12}, Coxeter–Todd lattice (see [CS88] with references). The minimum distance is $\xi(\boldsymbol{M}) = 2$, the determinant of the Gram matrix is $|\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}| = 729$, and the corresponding packing gain is $\sigma^2(\boldsymbol{M}) = 2.31$.

the fact that the volume of the Voronoi cells must be equal to the volume of the parallelotope defined by the columns of the generator matrix $\boldsymbol{M}_0$. The volume of the parallelotope is, in turn, equal to $|\boldsymbol{M}_0^{\mathrm{T}}\boldsymbol{M}_0|^{1/2}$. Because we assume $\boldsymbol{M}_0$ is $N \times N$ full rank, $|\boldsymbol{M}_0^{\mathrm{T}}\boldsymbol{M}_0|^{1/2} = |\boldsymbol{M}_0|$. The packing gain can be quite substantial for high dimensions N. Figure 5.2 shows an example of a 12 dimensional lattice with a substantial packing gain.

Applying the definition of packing gain (5.14) to (5.12) yields

$$P \geq \frac{N4^{R/N}}{16} \cdot \frac{\xi^2(\boldsymbol{M}_0)}{\sigma^2(\boldsymbol{M}_0)|\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H},N}|^{2/N}} = P_{LB}(\boldsymbol{M}_0), \tag{5.15}$$

which we refer to as the fundamental lower bound on the transmitted power. The bound depends to a large extent on the number of active subchannels, N. Weak subchannels will reduce the overall performance if activated. On the other hand, if N is too small, then too many bits are transmitted per subchannel which will also degrade the performance. By assuming the bound (5.15) is tight, it can be used to approximate the optimal value of N with a relative ease. Note that the bound also depends on the packing gain of the lattice — which also can be allowed to vary with N. In Section 5.3, the impact of the packing gain on the performance is further discussed.

It is interesting to compare the fundamental lower bound to the concept of optimal shaping gain, as described in Forney et al. [For99]. By making the so-called continuous approximation, one can show that the power-optimal codebook can be approximated by an infinite codebook, uniformly distributed inside a hyper-sphere

with radius Ψ. The optimal power is

$$P_{sphere} = \frac{1}{2^R} \sum_i \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_i \approx \frac{2N}{2N+1} \Psi^2 \approx \Psi^2. \tag{5.16}$$

The radius is related to the codebook size by equating the volume of the hypersphere with the volume of the union of all Voronoi regions in the codebook, $\boldsymbol{x} = \boldsymbol{F}\boldsymbol{s}$, as

$$2^R |\boldsymbol{F}| \approx \left(\frac{2\pi e \, \Psi^2}{N} \right)^{N/2}. \tag{5.17}$$

Combining the equations we get

$$P_{sphere} \approx \frac{N \, 4^{R/N} |\boldsymbol{F}|^{2/N}}{2\pi e}, \tag{5.18}$$

and since

$$|\boldsymbol{F}|^{2/N} = \frac{\xi^2(\boldsymbol{M}_0)}{\sigma^2(\boldsymbol{M}_0) |\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H},N}|^{2/N}} \tag{5.19}$$

we conclude that the fundamental lower bound is almost equal to the approximate power expression for the optimal shaping gain. This is interesting since the optimal constellation-shaping region is elliptical, hence very difficult to implement in practice. If our algorithm can reduce the power close to the lower bound, then it also performs close to the optimal constellation shape as described by Forney. (Note that since the optimal shaping region is not spherical, the classical upper bound, 1.53 dB, on the shaping gain is no longer valid.) The next step is to derive an upper bound that allows us to determine if, and when, the lower bound (5.15) is tight.

5.2.2 Upper bound on the optimized power

The upper bound on the transmit power is derived under the assumption that the bit-loading step in the optimization procedure has been performed. It is summarized in the following theorem:

Theorem 5.2.1 *Assuming the bit-loading Algorithm 5.1.2 has been performed: The transmitted power can be upper bounded as*

$$\begin{aligned} P &< \frac{N \cdot 4^{-2/3}}{e \, \ln(4)} 4^{R/N} \frac{|\boldsymbol{M}_0|^{2/N}}{|\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H},N}|^{2/N}} \, D(\boldsymbol{F}) \\ &< 1.69 \, P_{LB}(\boldsymbol{M}_0) \, D(\boldsymbol{F}) = P_{UB}(\boldsymbol{M}_0). \end{aligned} \tag{5.20}$$

Proof: Consider the transmit-signal basis matrix, $\boldsymbol{G}$, defined in (5.5), and denote the diagonal elements of $\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}$ as $g_1, ..., g_N$. The transmitted power can be upper bounded as

$$P = \frac{\sum_i g_i (4^{b_i} - 1)}{12} \leq \frac{\sum_i g_i 4^{b_i}}{12}. \tag{5.21}$$

Due to Algorithm 5.1.2, we know that there is at most a factor of 4 differing the maximum to the minimum term $g_1 4^{b_1}, ..., g_N 4^{b_N}$, in the summation. Applying Lemma 5.2.2 (stated below) with $C = 4$, we obtain

$$\sum_i g_i 4^{b_i} < \frac{3N \cdot 4^{1/3}}{e \ln(4)} \prod_i (g_i 4^{b_i})^{1/N}. \tag{5.22}$$

Hence,

$$P < \frac{N \cdot 4^{-2/3}}{e \ln(4)} 4^{R/N} |\boldsymbol{G}^{\mathrm{T}} \boldsymbol{G}|^{1/N} \, D(\boldsymbol{G}), \tag{5.23}$$

where we use the definition of $D(\cdot)$ from (5.8) to see that $D(\boldsymbol{F}) = D(\boldsymbol{G})$. The determinant is given by $|\boldsymbol{G}^{\mathrm{T}} \boldsymbol{G}| = |\boldsymbol{M}_0|^2 |\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H},N}|^{-2}$, which completes the proof. □

Lemma 5.2.2 *Let $\alpha_1, ..., \alpha_N$ be strictly positive real numbers, and let*

$$C = \frac{\max_i \alpha_i}{\min_j \alpha_j}, \tag{5.24}$$

then

$$f(\alpha_1, ..., \alpha_N) = \frac{\frac{1}{N} \sum_i \alpha_i}{\prod_i \alpha_i^{1/N}} < \frac{(C-1) C^{\frac{1}{C-1}}}{\ln(C)\, e}. \tag{5.25}$$

Proof: See Appendix 5.B. □

As we can see from (5.20), the upper bound and the lower bound differs by a factor $1.69\, D(\boldsymbol{F})$. Thus, in order to obtain an optimized power that is close to the fundamental lower bound, we need to make sure $D(\boldsymbol{F})$ is as close to one as possible. The upper bound (5.9) on $D(\boldsymbol{F})$, is exponentially increasing with the dimension N, but fortunately not very tight in most cases as is demonstrated in Section 5.4 where simulations of the CDF of $D(\boldsymbol{F})$ are presented for a MIMO link with 12 dimensions. However, if the minimized power in the end turns out to be far grater than the fundamental lower bound, then we can reduce the gap between the upper and lower bounds by enforcing a reduction of $D(\boldsymbol{F})$ through the choice of lattice base, $\boldsymbol{M}_0$.

5.3 Selecting the lattice base

In Section 5.1, it was demonstrated how to minimize the transmitted power for some fixed lattice base, $\boldsymbol{M}_0$. The next step is to determine what lattice base to use: The lower and the upper bounds from Chapter 5.2 on the transmitted power can facilitate this decision. As the first priority, our focus will be on finding a lattice base that minimizes the lower bound (5.15). This corresponds to finding the generator matrix, $\boldsymbol{M}_0$, of a certain dimension N that has the largest possible

Lattice	Dimension	$\sigma^2(\boldsymbol{M}_0)$ [dB]	$K(\boldsymbol{M}_0)$
$\mathbb{Z}$	1	0.000	2
A_2	2	0.625	6
A_3	3	1.003	12
D_4	4	1.505	20
D_5	5	1.806	40
E_6	6	2.215	72
E_7	7	2.580	126
E_8	8	3.010	240
Λ_9	9	3.010	272
Λ_{10}	10	3.135	336
K_{11}	11	3.305	432
K_{12}	12	3.635	756
K_{13}	13	3.722	918
Λ_{14}	14	3.960	1422
Λ_{15}	15	4.214	2340
Λ_{16}	16	4.515	4320
Λ_{17}	17	4.604	5346
Λ_{18}	18	4.752	7398
Λ_{19}	19	4.912	10668
Λ_{20}	20	5.118	17400
Λ_{21}	21	5.304	27720
Λ_{22}	22	5.530	49896
Λ_{23}	23	5.759	93150
Λ_{24}	24	6.021	196560

Table 5.1: Packing gains, $\sigma^2(\boldsymbol{M}_0)$, and kissing numbers, $K(\boldsymbol{M}_0)$, for lattices with dense packings in various dimensions.

packing gain, $\sigma^2(\boldsymbol{M}_0)$. The lattice sphere-packing problem is a classical problem in mathematics that has no known optimal solutions for most dimensions. Fortunately, many lattices with good sphere-packing properties have been found [CS88], and the best known lattice can be chosen for each dimensional size of interest. In Table 5.1, the packing gain, and the kissing number are listed for some well known lattices (that are available in e.g. [CS88]). If the minimized power is not sufficiently close to the lower bound after optimization, we can use the upper bound as described in Section 5.2.2 to understand how to search for lattice bases that are better.

5.3.1 Enforced reduction of the $D(\boldsymbol{F})$ factor

In the special case when the lattice equals $\mathbb{Z}^N$, all subchannels are orthogonal and $D(\boldsymbol{F}) = 1$. The corresponding upper bound is

$$P_{UB}(\boldsymbol{I}_N \xi) = 1.69\, P_{LB}(\boldsymbol{I}_N \xi). \tag{5.26}$$

For an arbitrary matrix lattice base $\boldsymbol{M}_0$ with dimension N and minimum distance ξ, the following relation holds

$$P_{LB}(\boldsymbol{I}_N\,\xi) = \sigma^2(\boldsymbol{M}_0)\,P_{LB}(\boldsymbol{M}_0). \tag{5.27}$$

Hence, if the minimized power P for lattice base $\boldsymbol{M}_0$, is a factor $1.69\,\sigma^2(\boldsymbol{M}_0)$ times larger than $P_{LB}(\boldsymbol{M}_0)$, then $P \geq P_{UB}(\boldsymbol{I}_N\,\xi)$, and it is always better to use the $\mathbb{Z}^N$-lattice rather than $\boldsymbol{M}_0$. The drawback is of course that the $\mathbb{Z}^N$-lattice does not provide any packing gain. So, a better strategy is to run the optimization with increasing orthogonalization — i.e. to force the lattice base to be more and more orthogonal — until a minimum in the transmitted power has been reached.

Let the lattice base, $\boldsymbol{M}_0$, be a concatenation of S sub-lattices as follows

$$\boldsymbol{M}_0 = \begin{bmatrix} \boldsymbol{M}_{0,1} & & \\ & \ddots & \\ & & \boldsymbol{M}_{0,S} \end{bmatrix}, \tag{5.28}$$

where $\boldsymbol{M}_{0,1}, ..., \boldsymbol{M}_{0,S}$ have dimensions $N_1, ..., N_S$, and sphere-packing gains $\sigma_1, ..., \sigma_S$. Due to orthogonality, the minimum distance of $\boldsymbol{M}_0$, is the minimum of the minimum distances of $\boldsymbol{M}_{0,1}, ..., \boldsymbol{M}_{0,S}$. Hence, it is power optimal to scale the sub-lattices to have the same minimum distance $\xi(\boldsymbol{M}_{0,i}) = \xi(\boldsymbol{M}_0)$: We assume that this scaling is applied. Due to the orthogonal subspaces in the lattice base, the power optimization algorithm will result in a block diagonal matrix

$$\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F} = \begin{bmatrix} \boldsymbol{F}_1^{\mathrm{T}}\boldsymbol{F}_1 & & \\ & \ddots & \\ & & \boldsymbol{F}_S^{\mathrm{T}}\boldsymbol{F}_S \end{bmatrix}. \tag{5.29}$$

We can now use the orthogonalization of the precoder to tighten the upper bound on $D(\boldsymbol{F})$ as

$$\begin{aligned} D(\boldsymbol{F}) &= \frac{\prod_{s=1}^{S}\prod_{i=1}^{N_s}[\boldsymbol{F}_s^{\mathrm{T}}\boldsymbol{F}_s]_{i,i}^{1/N}}{|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}|^{1/N}} \\ &= \prod_{s=1}^{S}\frac{\prod_{i=1}^{N_s}[\boldsymbol{F}_s^{\mathrm{T}}\boldsymbol{F}_s]_{i,i}^{1/N}}{|\boldsymbol{F}_s^{\mathrm{T}}\boldsymbol{F}_s|^{1/N}} \\ &\leq \prod_{s=1}^{S} 2^{\frac{N_s^2-N_s}{4N}} \leq 2^{\frac{N-S}{4}}, \end{aligned} \tag{5.30}$$

where we used (5.9) on each factor s. We conclude that by enforcing orthogonal subspaces, the upper bound on the orthogonalization factor can be reduced.

However, orthogonalization also affects the packing gain, which for the concatenated lattice can be calculated as the weighted geometric mean of the packing gains

of the sublattices

$$\sigma(\boldsymbol{M}_0) = \frac{\xi}{|\boldsymbol{M}_0|^{1/N}} = \prod_{s=1}^{S} \frac{\xi^{N_s/N}}{|\boldsymbol{M}_{0,s}|^{1/N}} = \prod_{s=1}^{S} \sigma(\boldsymbol{M}_{0,s})^{N_s/N}. \tag{5.31}$$

Because the highest possible packing gain for a sublattice (with dimension lower than N) is strictly lower than the highest possible packing gain of lattices of dimension N, we see that the orthogonalization reduces the packing gain. The price of introducing orthogonal subspaces is consequently an increased lower bound. A strategy is to let S go from 1 to N, keeping $N_1, ..., N_S$ as equal as possible, and then terminate when the power has reached a minimum. In the worst case, this would lead to the $\mathbb{Z}^N$-lattice, i.e. completely orthogonal transmission.

5.3.2 Kissing number versus packing gain

One conclusion from Section 5.3.1 is that, since the algorithm from Section 5.1.4 may not always be able to reach close to the lower bound, maximizing packing gain is not always the best choice. Moreover, there is also another negative effect: The packing gain is connected to the kissing number and, roughly speaking, the higher packing gain the larger is the kissing number. From equation (4.17), we observe that the kissing number has a direct negative impact on the performance, while (due to the relation with packing gain) it also has an indirect positive effect on the performance. There is no exact way to determine which effect that dominates from case to case. In the following, we present an approximation to set a heuristic rule for this tradeoff. By applying the Chernoff upper bound approximation (assuming high SNR) on the Q-function in the BLER expression (4.17), we have

$$P_e \approx 0.5\, K(\boldsymbol{M}) \exp\Big(- \frac{\xi^2(\boldsymbol{M})}{4} \Big). \tag{5.32}$$

Assume that the fundamental lower bound (5.15) is tight, such that the minimum distance can be approximated as

$$\xi^2(\boldsymbol{M}) \approx \frac{16P|\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H},N}|^{2/N}}{N4^{R/N}} \sigma^2(\boldsymbol{M}), \tag{5.33}$$

and then inserted into the BLER approximation. We now seek the lattice base that minimizes

$$P_e \approx 0.5\, K(\boldsymbol{M}) \exp\left(- \frac{4P|\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H},N}|^{2/N}}{N4^{R/N}} \sigma^2(\boldsymbol{M}) \right). \tag{5.34}$$

We have observed (empirically) that the kissing numbers for dense lattices roughly grow exponentially with increasing packing gain. Hence, we introduce the following approximation

$$K(\boldsymbol{M}) \approx N\, K_0 \exp\big(\gamma \sigma^2(\boldsymbol{M})\big), \tag{5.35}$$

where $K_0 = 0.1173$ and $\gamma = 2.751$. In Figure 5.3 the approximation is evaluated as a function of the true kissing numbers for the lattices in Table 5.1; we conclude that the fit is satisfactory. Consequently, the BLER may be approximated as

$$P_e \approx 0.5\, N\, K_0 \exp\left(\left(\gamma - \frac{4P|\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H},N}|^{2/N}}{N4^{R/N}}\right)\sigma^2(\boldsymbol{M})\right), \tag{5.36}$$

and thus, for low transmit powers an increased packing gain will actually result in a decreased performance. Using equation (5.36), we formulate the following rule of thumb: If the transmit power satisfies

$$P \gg \frac{0.7N\, 4^{R/N}}{|\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H},N}|^{2/N}}, \tag{5.37}$$

then the lattice base should have as large packing gain as possible. If the inequality does not hold, it is better to use lattices with low packing gain — for instance the $\mathbb{Z}^N$-lattice. This rule of thumb is based on approximations and some precaution must be taken. When $P \approx \frac{0.7N\, 4^{R/N}}{|\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H},N}|^{2/N}}$, the SNR is typically too low for the BLER approximation (5.32) to be tight. However, the general trend holds: The packing gain is more important in the high SNR region rather than the low SNR region.

5.3.3 Selection procedure

Below, step by step instructions are given to solve the joint bit loading and precoding problem for ML detection. Although some steps in the instructions are heuristic, they are all motivated by the theoretical considerations presented in this part of the theis.

1. The first step is to determine the maximum tolerable kissing number as specified by system constraints. The kissing number is, for instance, restricted by the largest possible lattice dimension — which equals the rank, Q, of the channel matrix. Complexity issues also pose constraints on the kissing number: The more nearest neighbors, the more complex is the decoding problem at the receiver. In this work, the maximum kissing number is regarded as a design parameter (that will not be optimized).

2. Determine the number of data subchannels, N, that should be used. This is achieved by minimizing the fundamental lower bound (5.15). When doing so, it is important to note that the diagonal matrix $\tilde{\boldsymbol{\Lambda}}_{\boldsymbol{H},N}$ also depends on N. For each N, we should use the lattice with the best possible sphere-packing density, and that has a kissing number lower than or equal to the maximum kissing number. At this point we have obtained a candidate for the lattice base $\boldsymbol{M}_0$.

3. Check the packing gain versus kissing number tradeoff. If the SNR is too low, i.e. if the rule of thumb (5.37) does not hold: Then the lattice base generator

matrix should be changed to $\boldsymbol{M}_0 = \boldsymbol{I}_N \xi$, in order to keep the kissing number small.

4. Optimize the power using the minimization algorithm in Section 5 for the candidate lattice base $\boldsymbol{M}_0$. If some subchannels have been turned off during the bit loading, reduce N by one, change the lattice base to the best lattice (in terms of packing gain) with this new dimension, and go back to step 4.

5. If the precoder has a $D(\boldsymbol{F})$ factor that is significantly larger than one (say by more than 2 dB), then the lattice basis reduction algorithm has not been able to orthogonalize the precoder sufficiently. In this case there is a potential gain by constructing a lattice base with concatenated lattices of lower dimension. Start with a lattice using two orthogonal subspaces of similar dimensions and go back to step 3. If the minimized power is reduced use the concatenated lattice. Iteratively test lattice bases with more and more orthogonal subspaces until the $D(\boldsymbol{F})$ factor falls below the threshold level (for example 2 dB). The threshold level is a design parameter which involves considerations regarding computational complexity.

The complexity of the transmission scheme depends, to a large extent, on the number of iterations in the lattice precoding algorithm. Fortunately, only a few iteration rounds are typically needed since both $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ are discrete, so that once they stabilize, the optimization of $\boldsymbol{U}$ needs only one last round.

5.4 Numerical results

The proposed scheme can provide a packing gain of several dB for sufficiently high SNR and sufficiently large lattice dimensions (compared to completely orthogonal transmission). Unfortunately packing gain is not the only factor affecting the performance: The kissing number can have a dominant negative effect for low to moderately low SNR. Furthermore, in order to fully benefit from the packing gain it is necessary that the basis reduction step (Section 5.1.3) manages to sufficiently orthogonalize the channel. It remains to be seen how much of the packing gain that can be realized in practice. In this section, we seek to shed some light on these issues through numerical simulations.

As was mentioned in Section 2.1, there are many potential applications for the proposed transmission scheme. For simplicity the examples herein are limited to narrow band multi-antenna systems of various dimensions, where the channel matrix consists of Rayleigh-fading matrix elements. This channel model is widely used for modelling MIMO channels with rich scattering around both the transmitter and the receiver. When the TX-CSI is perfect, as in this case, the definition of the SNR is not straightforward because the received signal power depends on the eigenvectors of the precoder. In the examples below, we avoid this problem by simply defining SNR as the transmitted power, P, (note that the noise power is normalized to one).

The proposed precoding scheme will be compared with three commonly used transmission schemes. The following list specifies the schemes:

- *Proposed precoding scheme*
  This is the bitloading and precoding scheme described in Section 5.3.3. The lattice with the best known sphere packing for the dimension of interest is used. These lattices are listed in Table 5.1 for all dimensions up to $N = 24$. After optimization the precoding matrices are re-scaled to ensure a specific transmit power, and consequently, the BLER is minimized instead of the power.

- *Orthogonalizing bit and power loading scheme*
  By running the power minimization algorithm on the $\mathbb{Z}^N$ lattice, we get the orthogonalizing precoder that after re-scaling maximizes the minimum-distance. In the high SNR region, this is the optimal orthogonalizing precoder. Note that this precoder creates parallel non-interfering sub-channels and the optimal ML decoder can therefore be implemented using the linear ZF detector.

- *Blind transmission scheme*
  Blind transmission is a relatively simple transmission scheme that does not utilize any TX-CSI, where the bit load is as evenly distributed as possible, and the transmitted signal vector is

$$\boldsymbol{c} = \sqrt{\frac{P}{\mathrm{Tr}\{\boldsymbol{\Sigma}\}}}\,\boldsymbol{x}. \tag{5.38}$$

  ML detection of blindly transmitted data can be computationally demanding, especially when the channel matrix is close to rank deficient.

- *Decentralized detection scheme*
  In addition to the above linear precoders, Tomlinson-Harashima (TH) precoding for decentralized receivers will be simulated for comparison [FWLH02a]: No joint processing of the received signals are needed. Scaling and a modulus operation on each signal is all that is needed. In order to make the comparison fair for the TH precoder, we implemented adaptive bit and power loading on the subchannels.

Figure 5.4 shows a comparison of the average BLER performance of the $\mathbb{Z}^{12}$ and K_{12}, when applied to a 6×6 MIMO channel. One thousand channel matrix realizations where used, with the matrix elements drawn from an uncorrelated Rayleigh-fading distribution as

$$\mathrm{vec}\big(\bar{\boldsymbol{H}}\big) \sim CN\big(\boldsymbol{0}, \boldsymbol{I}\big). \tag{5.39}$$

Precoding was performed for each realization at different SNR levels, and with a fixed data rate. The average block error rate was calculated using Monte Carlo

simulations, in which ML-detection was performed using the finite-alphabet constellation algorithm from [DGC03]. The maximum dimension of the lattices is 12, although lower dimensions are used whenever it improves the fundamental lower bound (5.15). The data rate is 24 bits per channel use.

We observe that, due to the packing gain of the K_{12} lattice, a gain of approximately 2 dB is attained over the $\mathbb{Z}^{12}$ lattice in the high SNR region. Judging from the sphere-packing gains listed in Table 5.1, one may come to the conclusion that the gain of the K_{12} lattice over the $\mathbb{Z}^{12}$ lattice should be on the order of 3.6 dB instead of 2 dB. There are three factors that can explain why this is not the case:

1. The kissing number of the K_{12} lattice is larger than for the $\mathbb{Z}^{12}$ lattice. According to equation (4.17) this will result in a higher BLER for the K_{12}-lattice in relative terms.

2. While the $\mathbb{Z}^{12}$ lattice is completely orthogonal, resulting in the tightest possible upper bound, the $D(\boldsymbol{F})$ factor for the K_{12} lattice is larger than one and there may be a loss due to lack of orthogonalization.

3. The algorithm adaptively selects the number of subchannels, N, and the optimal number of active channels is not always the largest possible (in this case 12). From Table 5.1, we know that a lower lattice dimension results in a lower packing gain.

In order to verify whether these three factors are sufficient for explaining the difference between the ideal gain and the measured gain, we have roughly estimated their values from the simulation. The results are shown in Table 5.2. The relative difference in kissing numbers is approximately 24, which, by analyzing slope of the curve of interest in Figure 5.4, roughly translates into a 1 dB loss in the high SNR region. The average $D(\boldsymbol{F})$-factor of K_{12} was evaluated numerically to 0.528 dB, so a loss of approximately 0.5 dB could be expected. In Figure 5.5, the cumulative density function (CDF) of $D(\boldsymbol{F})$ is approximated using a histogram. Finally, in Table 5.3 the third factor is analyzed using the histogram of the number of active subchannels (for the channel realizations used in the examples). It turns out that in most cases only 10 subchannels are active, and the packing gain is on average 0.31 dB lower than if all 12 dimensions would have been used. Adding all effects together, we get a residual gain of $3.3 - 0.5 - 1$ dB ≈ 1.8 dB, which is more in line with what can be observed in Figure 5.4. While the above analysis is approximate, we have seen that the three factors are plausible candidates for explaining the observed differences.

Returning to Figure 5.4, we can see that blind transmission works (perhaps surprisingly) well for low SNR considering that no TX-CSI is used. This may be explained by the low number of nearest neighbors the blindly transmitted codewords typically have. Remember that the negative effect of nearest neighbors is most prominent in the low-SNR region. While blind transmission leaves all the processing to the receiver (ML detection of blindly transmitted data is in general

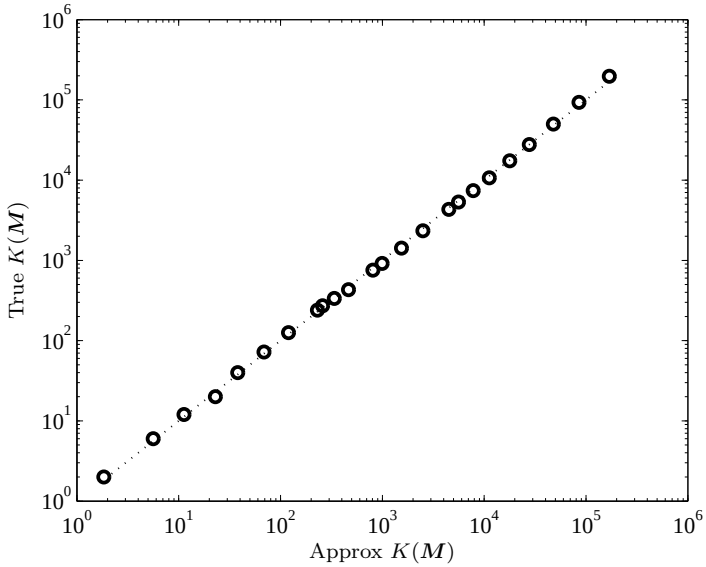

Figure 5.3: Evaluation of approximation (5.35) that approximates the kissing number as a function of the packing gain. The figure shows the approximated kissing number versus the true kissing number for the lattices in Table 5.1.

	K_{12}	$\mathbb{Z}^{12}$
Avg. Kissing Nbr.	488	20.6
Avg. $D(\boldsymbol{F})$	0.528 dB	0 dB
Avg. σ^2	3.324 dB	0 dB

Table 5.2: The average kissing numbers, sphere packing gains, and orthogonality factor for the lattice precoders simulated in Figure 5.4.

N	8	9	10	11	12	Total
$\mathbb{Z}^{12}$	6	7	798	65	124	1000
K_{12}	1	0	449	243	307	1000

Table 5.3: Histogram of the number of used dimensions, N, in the simulations from Figure 5.4.

computationally demanding), the TH precoder for decentralized receivers does the opposite. It has essentially all processing on the transmitter side, and the receiver can detect the data directly from the received signals, after a simple modulo operation. Although the TH precoder is well-suited for the case where the receiver antennas are decentralized, it has difficulties competing with the ML decoder that employs joint detection.

The conclusion from Figure 5.4 is: When using TX-CSI in combination with ML decoding, one needs to find a tradeoff between sphere-packing gain and kissing number. Interestingly, traditional channel diagonalization with adaptive bit loading (i.e. using $\mathbb{Z}^N$) does not meet this tradeoff, it can even perform worse than blind transmission. Another conclusion is that the detection algorithm plays a central role. In many cases blind transmission may perform excellently provided the appropriate detection algorithm is used.

These facts are even more evident for systems with higher dimensions. In Figure 5.6, an 8×8 MIMO channel with uncorrelated Rayleigh-fading elements is simulated in a similar fashion. The densest lattice for this scenario is the Λ_{16} lattice. With a data rate of 32 bits per channel use, the gain is on the order of 2–3 dB which is smaller than the packing gain of the lattice. The gap to the packing gain can again be explained by the three factors described above. For BLER levels of practical interest, the blind transmission scheme clearly outperforms $\mathbb{Z}^{16}$ lattice precoding. However, in the more realistic scenario with correlation between channel elements [YBO$^+$01], one can expect blind transmission to suffer a greater loss in performance compared to the $\mathbb{Z}^{16}$ precoder. This fact is illustrated in Figure 5.7, which shows the performance of the algorithm over an 8×8 correlated Rayleigh-fading channel. As opposed to the uncorrelated case, both lattice types now outperform blind transmission. In the simulations we used the so-called Kronecker correlation model [YBO$^+$01, SFGK00]: The covariance matrix is Kronecker structured, i.e.

$$\mathrm{E}\left\{\mathrm{vec}(\boldsymbol{H})\mathrm{vec}(\boldsymbol{H})^*\right\} = \boldsymbol{R}_t^{\mathrm{T}} \otimes \boldsymbol{R}_r, \tag{5.40}$$

with the transmit covariance matrix being parameterized as

$$[\boldsymbol{R}_t]_{ij} = \begin{cases} \rho_{\mathrm{t}}^{i-j} & i > j \\ 1 & i = j \\ \rho_{\mathrm{t}}^{*(j-i)} & i < j \end{cases}, \tag{5.41}$$

and with similar parametrization of $\boldsymbol{R}_r$ using ρ_{r}. The covariance coefficients $\rho_{\mathrm{t}}, \rho_{\mathrm{r}} \in \{z : |z| < 1\}$, represent the amount of transmit and receive covariance respectively. That is, the covariance decreases exponentially with the antenna-index distance. This model might be realistic for uniform linear arrays in rich scattering environments. For more motivations and references on this model see [MO04]. In the example herein $\rho_{\mathrm{r}} = \rho_{\mathrm{t}} = 0.6 + j0.1$ is used.

Although the packing gain can not be realized in its entirety, we have seen that it is possible to achieve a significant part of it. Furthermore, for uncorrelated

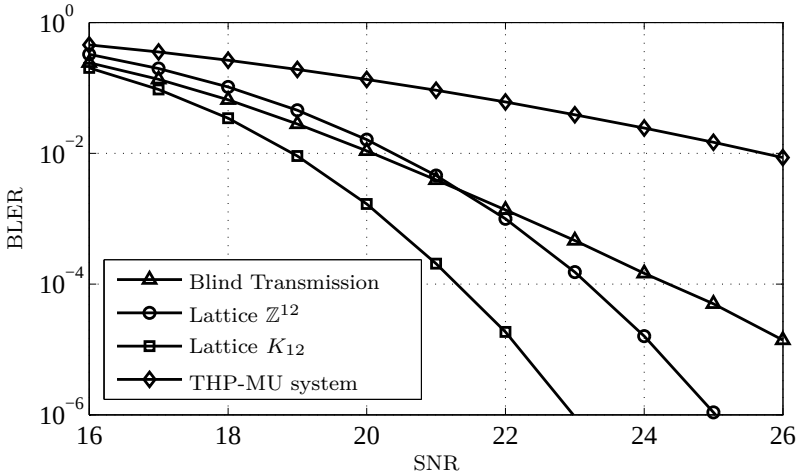

Figure 5.4: Block error rate comparison between the $\mathbb{Z}^{12}$ lattice, the K_{12} lattice, Tomlinson–Harashima precoding for decentralized receivers, and 12-dimensional blind transmission. The channel is a 6×6 MIMO Rayleigh-fading channel, and the data rate is 24 bits/use.

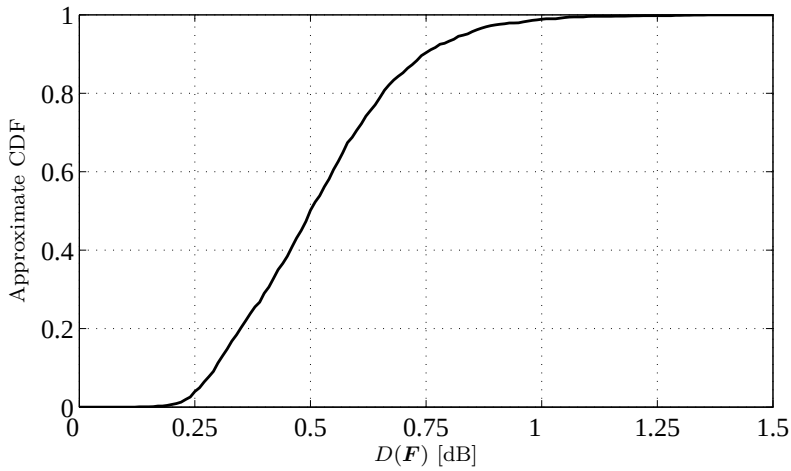

Figure 5.5: Monte Carlo approximation of the CDF of the loss due to the orthogonalization factor $D(\boldsymbol{F})$ for the K_{12} lattice in Figure 5.4.

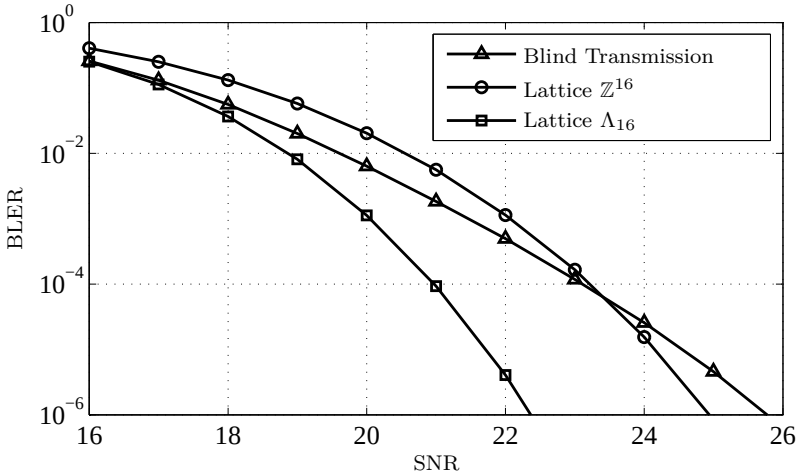

Figure 5.6: Block error rate comparison between the $\mathbb{Z}^{16}$ lattice, the Λ_{16} lattice, and 16-dimensional blind transmission. The channel is an 8×8 MIMO Rayleigh-fading channel, and the data rate is 32 bits/use.

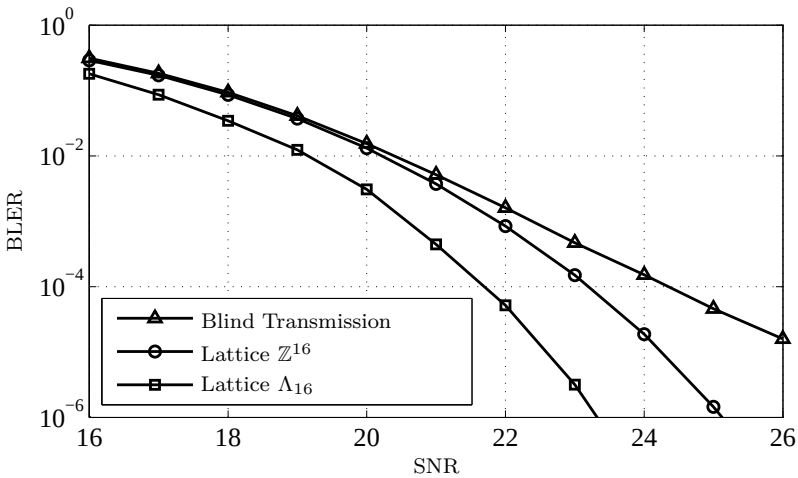

Figure 5.7: Block error rate comparison between the $\mathbb{Z}^{16}$ lattice, the Λ_{16} lattice, and 16-dimensional blind transmission. The channel is an 8×8 MIMO Rayleigh-correlated fading channel, and the data rate is 26 bits/use.

Rayleigh-fading channels, it is in fact necessary to use lattices with high packing gain in order to outperform blind transmission in terms of minimum BLER.

5.5 Conclusions to Part I

The problem of designing bit loading and linear precoder for MIMO communication systems has been investigated. Both the receiver and the transmitter are assumed to have perfect CSI; the receiver is assumed to employ ML detection. The main conclusion is that orthogonal transmission in combination with bit and power loading is not optimal if the receiver uses ML detection. In some cases even blindly transmitted data performs better than the orthogonal scheme. In order to fully take advantage of the ML decoder, we propose the use of lattice invariant operations to transform the channel matrix into a lattice generator matrix. The lattice should ideally have good sphere-packing properties, resulting in large minimum-distance separation. An algorithm that performs the transformation in a close to power-optimal manner was presented, along with a lower and an upper bound on the optimized power level. The lower bound motivates the use of lattice generator matrices with dense sphere-packing properties, although it was concluded that the kissing number of the lattice also affects the performance. A design methodology was presented using the theoretical developments herein. Numerical results indicate that there is a potential gain of several dB by using the method compared to channel diagonalization with adaptive bit loading.

Appendix 5.A Proof of Theorem 5.1.2

Let $\bar{b}_1, ..., \bar{b}_N$ be a local optimum, i.e satisfying

$$g_n 4^{\bar{b}_n} \leq 4\, g_m 4^{\bar{b}_m} \tag{5.42}$$

for all n and m such that $\bar{b}_n > 0$. Let $b_1, ..., b_N$ be an arbitrary bit load. Define $\delta_n = b_n - \bar{b}_n$ for all n. Note that all bit load residuals $\delta_1, ..., \delta_N$ are integer-valued. Let the transmitted power for the local optimum be $\bar{P}$ and the transmitted power for $b_1, ..., b_N$ be P. We have the difference in power

$$\begin{aligned} 12(P - \bar{P}) &= \sum_i g_i 4^{\bar{b}_i}(4^{\delta_i} - 1) \\ &= \sum_{i:\delta_i > 0} g_i 4^{\bar{b}_i}(4^{\delta_i} - 1) + \sum_{i:\delta_i < 0} g_i 4^{\bar{b}_i}(4^{\delta_i} - 1). \end{aligned} \tag{5.43}$$

Define

$$C = \min_{i:\delta_i > 0} g_i 4^{\bar{b}_i}, \tag{5.44}$$

so that we can bound

$$\sum_{i:\delta_i > 0} g_i 4^{\bar{b}_i}(4^{\delta_i} - 1) \geq C \sum_{i:\delta_i > 0} (4^{\delta_i} - 1). \tag{5.45}$$

For all $\delta_n < 0$ we know that $\bar{b}_n > 0$ and consequently by using local optimality $g_n 4^{\bar{b}_n} \leq 4C$. Assembling these results we have

$$12(P - \bar{P})/C \geq \sum_{i:\delta_i > 0} (4^{\delta_i} - 1) + 4 \sum_{i:\delta_i < 0} (4^{\delta_i} - 1). \tag{5.46}$$

Let M_1 be the number of elements in $\delta_1, ..., \delta_N$ such that $\delta_i = -1$. Let M_2 be the number of elements that satisfy $\delta_i < -1$. It is easily verified that

$$4 \sum_{i:\delta_i < 0} (4^{\delta_i} - 1) \geq -3\, M_1 - 4\, M_2. \tag{5.47}$$

Similarly it can be shown that for integer $\delta_i > 0$ we have

$$\sum_{i:\delta_i > 0} (4^{\delta_i} - 1) \geq 3 \sum_{i:\delta_i > 0} \delta_i. \tag{5.48}$$

Since the total data rate is the same for both bit loads, $b_1, ..., b_N$ and $\bar{b}_1, ..., \bar{b}_N$, we know that $\sum_i \delta_i = 0$ and consequently

$$\sum_{i:\delta_i > 0} \delta_i = \sum_{i:\delta_i < 0} -\delta_i \geq M_1 + 2M_2. \tag{5.49}$$

Assembling the results we have

$$\begin{aligned} 12(P - \bar{P})/C &\geq 3(M_1 + 2M_2) - 3\, M_1 - 4\, M_2 \\ &= 2M_2 \geq 0. \end{aligned}$$

Hence $P \geq \bar{P}$, and a local optimum is also a global optimum. □

Appendix 5.B Proof of Lemma 5.2.2

We are seeking the maximum of

$$f(\alpha_1, ..., \alpha_N) = \frac{\frac{1}{N}\sum_{i=1}^{N} \alpha_i}{\prod_{j=1}^{N} \alpha_j^{1/N}}, \tag{5.50}$$

which is equivalent to maximizing $g(\cdot) \triangleq \log f(\cdot)$. Normalize the variables as

$$\beta_i = \frac{\alpha_i}{\min_j \alpha_j}, \tag{5.51}$$

such that $1 \leq \beta_i \leq C$ for all i, where

$$C = \frac{\max_i \alpha_i}{\min_j \alpha_j}. \tag{5.52}$$

Note that the normalization does not affect the cost function,

$$g(\beta_1, ..., \beta_N) = g(\alpha_1, ..., \alpha_N). \tag{5.53}$$

Assume that N_1 variables are forced to satisfy $\beta_i = 1$, and N_C variables satisfy $\beta_i = C$. Differentiating $g(\cdot)$ with respect to the remaining variables $\beta_1, ..., \beta_{N-N_1-N_C}$ and equating to zero results in the following system of equations

$$\frac{1}{N_1 + N_C C + \sum_{i=1}^{N-N_1-N_C} \beta_i} = \frac{1}{N\beta_j} \quad \forall\, j = 1, ..., N - N_1 - N_C, \tag{5.54}$$

with the solution $\beta_i = \frac{N_1 + N_C C}{N_1 + N_C}$ for all i. The Hessian at this optimal point is always positive semi-definite, hence the optimum is a minimum. Any maximum must consequently have some of $\beta_1, ..., \beta_{N-N_1-N_C}$ equal to either 1 or C. By recursion, the maximum of $f(\beta_1, ..., \beta_N)$ must have some N_1 elements equal to 1 and $N - N_1$ elements equal to C. Then we have

$$f(1, ..., 1, C, ..., C) = \frac{\frac{N_1}{N} + \frac{N-N_1}{N} C}{C^{\frac{N-N_1}{N}}} = \frac{1 + (C-1)\gamma}{C^{\gamma}}, \tag{5.55}$$

where $\gamma = \frac{N-N_1}{N}$. Maximizing with respect to γ gives

$$\gamma^* = \frac{1}{\ln C} - \frac{1}{C-1}, \tag{5.56}$$

and the resulting upper bound is (note that for $C > 1$ this gives a valid $\gamma^* \in (0, 1)$)

$$f(\cdot) < \frac{(C-1)C^{\frac{1}{C-1}}}{\ln(C)\, e}. \tag{5.57}$$

□

Part II

Design based on decision feedback detection

Chapter 6

Introduction to Part II

Under the assumption that the transmitter knows the channel perfectly, the capacity-optimal precoding strategy is to linearly orthogonalize the channel matrix using the SVD [CT91, Tel95]. Information is optimally conveyed over the orthogonal subchannels using infinitely long and Gaussian distributed codewords, with data rates assigned to the subchannels given by the so called waterfilling solution. Although SVD based, orthogonal, transmission is optimal in the sense of maximizing the mutual information, it is not necessarily optimal in the delay-limited case that is considered herein. In Part I of this thesis, the suboptimality of orthogonal transmission was shown given that the optimal ML decoder is used at the receiver. Another design was proposed in [CBRB04] that gives the minimum BER solution for a 2×2 MIMO system with quadrature phase-shift keying modulation. Both of these designs include a rotation of the precoder such that the effective channel is not orthogonalized.

The ML detection problem can be rather computationally demanding if the channel is not orthogonal, and therefore suboptimal receivers are often considered. Arguably one of the more commonly considered receiver structures is the linear detector (ZF or MMSE). When using equal bit rates on all active subchannels, the optimal precoding strategy for linear receivers was given in [PCL03, DDLW03]. It turns out that the best choice in terms of minimum BER, is to mix the substreams using a discrete fourier transform (DFT) matrix so that all substreams are dispersed equally over all channel eigenvalues. In [PCL03], the precoding problem for linear receivers was generalized to cost functions of the mean squared errors (MSE) that are either Schur-convex or Schur-concave. A short definition of Schur-convexity and concavity is included in Appendix 9.A. For Schur-convex cost functions, the solution is similar to the solution with the minimum BER objective, i.e., it is optimal to mix the subchannels (in fact, for equal signal constellations the minimum BER objective is Schur-convex). For Schur-concave cost functions, the optimal solution is instead to transmit on orthogonal subchannels using SVD based precoding, similar to the capacity optimal transmission.

The analysis in [PCL03, DDLW03] did not consider using optimized bit loading on the subchannels. Adjusting the data rate according to the channel quality is important in order to achieve high performance in a digital communication system. In [DDW03, PLC04, PBO05] the problem of designing transmit and receive filters when using heterogenous constellations on the subchannels were treated. Different bit rates on the subchannels results in filters that need to meet various quality of service constraints depending on the constellations that are used. This question was taken one step further in [PB05] where the joint optimization of bit loading and linear precoder was investigated. Surprisingly, perhaps, it was shown that joint optimization of the bit loading, transmit and receive filters results in orthogonal subchannels and that no DFT-type of rotation should be used. In other words — the bit-loading optimized cost function is Schur-concave with respect to the MSEs.

To summarize; for joint bit loading and precoding it is known that orthogonal transmission is suboptimal for ML receivers (Part I), but optimal for linear receivers [PB05]. An intermediate receiver solution between the linear receiver and the ML detection is the DF receiver [BP79, GC01, WFGV98, Gue03, SGS01, XDZW06, JHL06, SD08, PJ07]. The DF receiver has low decoding complexity compared to the ML detector [JO05], and for channels with inter-symbol interference it outperforms the linear receiver in terms of error performance (the linear receiver is a special case of the DF receiver). Since the DF decoder is something in between the ML and the linear detector in terms of performance, a natural question to ask is whether the optimal joint bit loading and precoding strategy (when using DF) is to orthogonalize the channel or not? Note that if the precoder orthogonalizes the channel, the DF detector has no inter-symbol interference to remove, and thus, linear detection and DF detection becomes equivalent.

In [XDZW06, PJ07, SD08] it was shown that orthogonal transmission is indeed optimal for multiplicative Schur-concave objective functions of the MSEs, while for multiplicative Schur-convex objectives a rotation of the signal vector is needed (similar to the linear case). These conclusions are for systems with equal constellations, and as a relevant special case the minimum BER objective is shown to be multiplicatively Schur-convex. In [JHL06, PJ07] the quality-of-service constrained problem was treated, which allows for optimal filter design given a fixed heterogenous bit loading.

In this part of the thesis we consider the problem of jointly optimizing the precoder, receiver filters, and bit loading when using the DF receiver. The main result is that the optimal bit loading will, in fact, result in an orthogonalizing precoder design. Orthogonal subchannels implies that there is no interference for the DF detector to remove, and as a result decision feedback actually becomes superfluous when the signal constellations are chosen properly. However, due to robustness the DF may still be advantageous to implement: The DF allows us to redistribute the bit loading on high-rate subchannels at a very low cost in terms of reduced performance, which in turn means that a suboptimal bit loading will perform almost as good as the optimal one. Another reason for using DF detection is that perfect transmitter-side CSI may be an unrealistic assumption. Imperfect

transmitter-side CSI inevitably causes inter-symbol interference that can be reduced using DF (note that perfect CSI is required for orthogonal transmission).

In addition to the results regarding bit loading, we show that the problem of computing the optimal precoder and receiver filters for a fixed bit loading can be posed as a convex problem (similar, but not equivalent to [JHL06]). An algorithm that solves the convex problem with linear computational complexity is provided. Because of the low computational complexity of the filter optimization, an exhaustive search for the optimal bit loading becomes feasible in practice — although the main result of the paper suggests that an exhaustive search is not necessary in most cases.

Finally, as a byproduct from the work regarding linear precoding for DF receivers, we include Chapter 10 that treats a special class of optimization problems with so-called skewed majorization constraints. It turns out that problems of this class are particularly easy to solve by simply identifying the convex hull under a sequence of numbers. Two MIMO related problems are provided as applications to this class of problems.

Chapter 7

Performance measure and problem formulation

Part II of the thesis considers joint bit loading and linear precoding assuming delay-limited transmission and a minimum-MSE decision feedback detector at the receiver. In this chapter we derive a performance measure and define the optimization problem for this particular setting.

7.1 System model

Consider the discrete-time flat-fading linear model of a $N_r \times N_t$ MIMO communication system

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{F}\boldsymbol{s} + \boldsymbol{n}, \tag{7.1}$$

where $\boldsymbol{y} \in \mathbb{C}^{N_r}$ is the received signal, $\boldsymbol{H} \in \mathbb{C}^{N_r \times N_t}$ is the channel matrix, $\boldsymbol{F} \in \mathbb{C}^{N_t \times N}$ is a precoding matrix, $\boldsymbol{s} \in \mathbb{C}^{N}$ is the data-symbols vector, and $\boldsymbol{n} \in \mathbb{C}^{N_r}$ is additive white circularly-symmetric complex-Gaussian noise. The data symbols and the noise are assumed to be normalized as $\mathrm{E}[\boldsymbol{s}\boldsymbol{s}^*] = \boldsymbol{I}$ and $\mathrm{E}[\boldsymbol{n}\boldsymbol{n}^*] = \boldsymbol{I}$. The average transmitted power is limited such that $\mathrm{Tr}\{\boldsymbol{F}\boldsymbol{F}^*\} \leq P$, is satisfied.

7.2 Decision feedback receiver

In the following chapters we assume that the receiver employs DF equalization, and a schematic view of the considered system is depicted in Figure 7.1. The received signal is linearly equalized using a forward filter, $\boldsymbol{W}^*$, and subsequently passed to an elementwise detector of the data symbols. From the outcome of the detection, we reconstruct the transmitted data symbols, $\tilde{\boldsymbol{s}}$, then use the reconstructed symbols to remove inter-symbol interference between the symbols in the equalized signal. In order to ensure that the DF detection is sequential, i.e. that we do not feedback symbols that have not yet been detected, we enforce the feedback matrix $\boldsymbol{B}$ to be

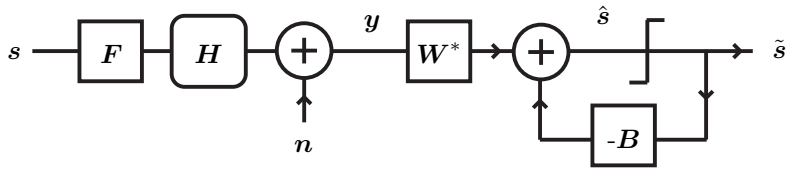

Figure 7.1: Schematic view of the MIMO communication system with DF detection.

strictly lower triangular. The signal after the interference subtraction, $\hat{\boldsymbol{s}}$, is then passed on to the detector again. Taking the feedback into account, the error prior detection is

$$\boldsymbol{e} = \hat{\boldsymbol{s}} - \boldsymbol{s} = (\boldsymbol{W}^*\boldsymbol{H}\boldsymbol{F} - \boldsymbol{I})\boldsymbol{s} - \boldsymbol{B}\tilde{\boldsymbol{s}} + \boldsymbol{W}^*\boldsymbol{n}. \tag{7.2}$$

If the probability of detection error is small, one can assume that $\tilde{\boldsymbol{s}}$ and $\boldsymbol{s}$ are zero mean and have almost identical auto-correlation and cross-correlation matrices, and that $\tilde{\boldsymbol{s}}$ is uncorrelated with the noise $\boldsymbol{n}$. Using these approximations the error covariance matrix is given by

$$\begin{aligned} \boldsymbol{R}_{\mathrm{MSE}} &= \mathrm{E}\left[\boldsymbol{e}\boldsymbol{e}^*\right] \\ &= (\boldsymbol{W}^*\boldsymbol{H}\boldsymbol{F} - \boldsymbol{B} - \boldsymbol{I})(\boldsymbol{W}^*\boldsymbol{H}\boldsymbol{F} - \boldsymbol{B} - \boldsymbol{I})^* + \boldsymbol{W}^*\boldsymbol{W}. \end{aligned} \tag{7.3}$$

Since detection of the symbols $\hat{\boldsymbol{s}} = \boldsymbol{s} + \boldsymbol{e}$ is made elementwise, we can regard the problem simply as detecting a scalar signal in additive complex-Gaussian noise[1]. We denote each virtual transfer function $\hat{s}_i = s_i + e_i$ as a subchannel, for which the performance is determined by its virtual noise power, $[\boldsymbol{R}_{\mathrm{MSE}}]_{i,i}$.

In general, sequential decision feedback does not restrict us to use only lower-triangular feedback matrices, any joint row–column permutation is also possible. However, in this case where we are free to design both the precoder and the bit loading, such a permutation loses its purpose since it can be absorbed into the other optimization parameters.

7.3 Cost functions based on the weighted mean squared error

A general framework was presented in [PJ07] for optimizing the DF system (i.e. the filters $\boldsymbol{F}$, $\boldsymbol{W}^*$, and $\boldsymbol{B}$) based on monotonic cost functions of the MSEs of the subchannels. Our goal here is to optimize not only these DF filters, but also the signal constellations that are used on the subchannels. For mathematical tractability in the later analysis, we narrow down the class of cost functions to p-norms of weighted

[1]Strictly speaking the interference part of the error is not complex Gaussian distributed. However, combined with the noise we can tightly approximate the interference as such by the law of large numbers.

MSEs. More precisely, consider the cost function $\|\boldsymbol{d}(\boldsymbol{R}_{\text{MSE}}\boldsymbol{D}_w)\|_p$, where $\boldsymbol{D}_w$ is a weighting matrix assumed to be diagonal and non-negative, and the function

$$\|\boldsymbol{d}(\boldsymbol{X})\|_p = \left(\sum_{i=1}^{N} [\boldsymbol{X}]_{i,i}^p\right)^{1/p}, \tag{7.4}$$

is the p-norm of the diagonal elements of $\boldsymbol{X}$. The p-norm is defined for $p \geq 1$.

To illustrate how the cost function (7.4) can be applied in practice, consider minimizing the probability of detection error. Using the Gaussian-tail function (2.43), the probability of error of subchannel i can be approximated as

$$P_{e,i} \simeq 4\,Q\left(\sqrt{\frac{d_{\min}^2(b_i)}{2\,[\boldsymbol{R}_{\text{MSE}}]_{i,i}}}\right), \tag{7.5}$$

where $d_{\min}^2(b_i)$ denotes the squared minimum distance of a b_i-bit signal constellation, normalized to unit variance [Pro01]. Equation (7.5) allows us to relate the MSE with the performance in terms of error probability. It also indicates how we should chose the MSE weighting matrix $\boldsymbol{D}_w$ in our cost function. Namely, in order to have symmetry among the subchannels, the weights should be inversely proportional to the squared minimum distance as

$$[\boldsymbol{D}_w]_{i,i} = d_{\min}^{-2}(b_i) \quad \forall\, i = 1, ..., N. \tag{7.6}$$

This will make the subchannels (approximately) symmetric with respect to SER, which is a relevant measure, for example, if we want to minimize the joint probability of detection error. In the case when outer error correcting codes are used it may be more relevant to have symmetric BERs rather than SERs. Assuming Gray coded bit mapping the BERs can be approximated as

$$\text{BER}_i \approx \frac{1}{b_i}\,P_{e,i}. \tag{7.7}$$

For moderately low BERs, it can be shown that the dependency on b_i in the BER expression is dominated by the SER factor, $P_{e,i}$. Hence symmetric SERs can serve as a good approximation to attain symmetric BERs as well.

For most classes of constellations used in practice, the minimum distance typically decreases exponentially with the number of bits b_i. For example, QAM constellations with even bit loading has minimum distance

$$d_{\min}^2(b_i) = \frac{6}{2^{b_i} - 1}, \tag{7.8}$$

resulting in a weighting matrix (disregarding constant factors)

$$\boldsymbol{D}_w = \boldsymbol{D}(2^{b_1} - 1, ..., 2^{b_N} - 1). \tag{7.9}$$

One objective could be to minimize the maximum error probability, $P_{e,i}$, of the subchannels. Under the high-SNR assumption, this objective translates into a cost function

$$\|\boldsymbol{d}(\boldsymbol{R}_{\mathrm{MSE}}\boldsymbol{D}_w)\|_\infty = \max_i [\boldsymbol{R}_{\mathrm{MSE}}\boldsymbol{D}_w]_{i,i}, \tag{7.10}$$

corresponding to the $p = \infty$ norm.

Another strategy is to have approximately equal error rate on all subchannels, but to allow small deviations around this point. To do this, we first apply the Chernoff upper bound [SA00] to approximate the Gaussian-tail function as

$$\sum_{i=1}^{N} P_{e,i} \approx \sum_{i=1}^{N} 2\exp\Big(-\frac{d_{\min}^2(b_i)}{4\,[\boldsymbol{R}_{\mathrm{MSE}}]_{i,i}}\Big). \tag{7.11}$$

The Taylor expansion of the Chernoff approximation around the point

$$[\bar{\boldsymbol{R}}_{\mathrm{MSE}}\boldsymbol{D}_w]_{i,i} = \kappa \quad \forall\, i = 1, ..., N, \tag{7.12}$$

where κ is a constant, gives

$$\sum_{i=1}^{N} P_{e,i} \approx 2N\Big(1 - \frac{6}{8\kappa}\Big)e^{-6/8\kappa} + \frac{12}{8\kappa^2}e^{-6/8\kappa}\,\mathrm{Tr}\{\boldsymbol{R}_{\mathrm{MSE}}\boldsymbol{D}_w\}. \tag{7.13}$$

Disregarding positive coefficients and constants, this cost function corresponds to the $p = 1$ norm

$$\|\boldsymbol{d}(\boldsymbol{R}_{\mathrm{MSE}}\boldsymbol{D}_w)\|_1 = \mathrm{Tr}\{\boldsymbol{R}_{\mathrm{MSE}}\boldsymbol{D}_w\}. \tag{7.14}$$

Summarizing, if the SNR is high, the probability of error on a subchannel depends to a large extent on the minimum distance of the signal constellations. The minimum distance scales the MSE of the subchannels, which leads to imbalances when different types of signal constellations are used on different subchannels. Using a cost function with weighted MSE this imbalance can be compensated for. The parameter p of the cost function can be used to control how *flexible* the system is in terms of the spread of the error rates among the subchannels. Low p results in more spread, which may be disadvantageous since the worst subchannel typically dominates. In general, and specifically for high SNRs, the infinity norm seems to translate into the lowest SERs in most cases.

7.4 Problem formulation

With the definitions of the MSE matrix (7.3) and the cost function (7.4) in place, our problem can be mathematically formulated as

$$\underset{\boldsymbol{F},\boldsymbol{B},\boldsymbol{W}^*,\boldsymbol{b}}{\text{minimize}} \quad \left\| \boldsymbol{d}\big(\boldsymbol{R}_{\text{MSE}}(\boldsymbol{F},\boldsymbol{B},\boldsymbol{W}^*)\boldsymbol{D}_w\big) \right\|_p \tag{7.15a}$$

$$\text{subject to} \quad \text{Tr}\{\boldsymbol{F}\boldsymbol{F}^*\} \leq P, \tag{7.15b}$$

$$[\boldsymbol{D}_w]_{i,i} = d_{\min}^{-2}(b_i) \quad \forall\, i = 1, ..., N, \tag{7.15c}$$

$$b_i \in \mathcal{B} \quad \forall\, i = 1, ..., N, \tag{7.15d}$$

$$\sum_{i=1}^{N} b_i = R, \tag{7.15e}$$

where the vector $\boldsymbol{b} = [b_1, ..., b_N]$ is the bit loading vector, and the set $\mathcal{B}$ denotes the set of feasible bit rates which is determined by the available signal constellations. Typically, due to the discrete nature of bits, this set is equal to the set of positive integers.

The problem of designing the DF filters for a fixed bit loading, $\boldsymbol{b}$, is treated in Chapter 8. It is shown how to apply the framework given in [PJ07] to the particular problem considered here, and the result is a problem formulation involving majorization inequality constraints. Then we show how the resulting non-convex problem can be replaced with a convex problem that can be solved very efficiently with linear complexity. Once the optimal bit rates are known, the remaining problem is therefore fairly simple.

As for the optimization of the bit loading, the set $\mathcal{B}$ is discrete and it is possible to numerically try out all feasible bit loading combinations in order to find the global optimum. An alternative to such an exhaustive search is to relax the problem and extend the set $\mathcal{B}$ to allow for arbitrary positive bit rates. We do this in Chapter 9 by using (7.8) to relate the real-valued bit rates to virtual minimum-distance weights. This relaxation allows us to optimize the bit loading for any given MSE matrix. The remaining problem of jointly optimizing the DF filters is characterized in Section 9.3, together with a discussion on the loss due to rounding of the bit rates. Finally, in Sections 9.5 and 9.6, various practical strategies for solving the joint problem are presented and evaluated numerically.

Chapter 8

Design of optimal DF filters

Because the set of feasible bit loads is discrete, its optimization is not easy to combine with the filter and precoder design. This chapter treats the filter design problem alone, assuming a fixed bit loading. The problem is formulated as

$$\underset{\boldsymbol{F},\boldsymbol{B},\boldsymbol{W}^*}{\text{minimize}} \quad \|\boldsymbol{d}(\boldsymbol{R}_{\text{MSE}}\boldsymbol{D}_w)\|_p \tag{8.1a}$$

$$\text{subject to} \quad \text{Tr}\{\boldsymbol{F}\boldsymbol{F}^*\} \leq P. \tag{8.1b}$$

In [PJ07, Theorem 4.3] it was shown how problems of this type (with monotonic cost functions of the MSEs) can be reduced to a power-loading problem involving a multiplicative majorization constraint[1]. In Sections 8.1, 8.2, and 8.3, we will, for completeness, apply this procedure to Problem (8.1), and derive expressions for the filters that are obtained in the process. In Section 8.4, we then show how the resulting problem with a majorization constraint can be replaced with a convex problem that is easy to solve numerically.

8.1 Optimal forward receiver filter

For a given transmit matrix, $\boldsymbol{F}$, and receive DF matrix, $\boldsymbol{B}$, the optimal forward filter, $\boldsymbol{W}^*$, in the receiver is the well known MMSE equalizer

$$\boldsymbol{W}^* = (\boldsymbol{B}+\boldsymbol{I})(\boldsymbol{F}^*\boldsymbol{H}^*\boldsymbol{H}\boldsymbol{F}+\boldsymbol{I})^{-1}\boldsymbol{F}^*\boldsymbol{H}^*. \tag{8.2}$$

The proof is given by completing the squares of (7.3) and then applying the matrix inversion lemma[2]. Using the optimal forward filter, the resulting vector of weighted MSEs is

$$\boldsymbol{d}(\boldsymbol{R}_{\text{MSE}}\boldsymbol{D}_w) = \boldsymbol{d}\big((\tilde{\boldsymbol{B}}+\boldsymbol{D}_w^{1/2})(\boldsymbol{F}^*\boldsymbol{H}^*\boldsymbol{H}\boldsymbol{F}+\boldsymbol{I})^{-1}(\tilde{\boldsymbol{B}}+\boldsymbol{D}_w^{1/2})^*\big), \tag{8.3}$$

[1] A short recapitulation on vector majorization is provided in Appendix 9.A.

[2] $(\boldsymbol{A}+\boldsymbol{B}\boldsymbol{C}\boldsymbol{D})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{B}+\boldsymbol{C}^{-1})^{-1}\boldsymbol{D}\boldsymbol{A}^{-1}$

where $\tilde{\boldsymbol{B}} \triangleq \boldsymbol{D}_w^{1/2}\boldsymbol{B}$. Note that the MMSE equalizer minimizes any monotonic increasing function of the MSEs and is thus optimal for a wider class of cost functions than considered here.

8.2 Optimal feedback receiver filter

Consider the Cholesky factorization of

$$(\boldsymbol{F}^*\boldsymbol{H}^*\boldsymbol{H}\boldsymbol{F} + \boldsymbol{I})^{-1} = \boldsymbol{L}\boldsymbol{L}^*, \tag{8.4}$$

where $\boldsymbol{L}$ is lower triangular. Inserting (8.4) into (8.3) we obtain the weighted MSE of subchannel i as

$$[\boldsymbol{R}_{\mathrm{MSE}}\boldsymbol{D}_w]_{i,i} = \sum_{j=1}^{i} |[(\tilde{\boldsymbol{B}} + \boldsymbol{D}_w^{1/2})\boldsymbol{L}]_{i,j}|^2. \tag{8.5}$$

Since $\tilde{\boldsymbol{B}}\boldsymbol{L}$ is a strictly lower-triangular matrix (zero diagonal) it can only affect the non-diagonal elements of the lower-triangular matrix $\tilde{\boldsymbol{B}}\boldsymbol{L} + \boldsymbol{D}_w^{1/2}\boldsymbol{L}$. Hence, the feedback matrix, $\boldsymbol{B}$, that minimizes the MSEs in (8.5) satisfies

$$\tilde{\boldsymbol{B}}\boldsymbol{L} = -\mathcal{L}_{\mathrm{strict}}\big(\boldsymbol{D}_w^{1/2}\boldsymbol{L}\big), \tag{8.6}$$

where $\mathcal{L}_{\mathrm{strict}}(\boldsymbol{X})$ is the matrix that contains the strictly lower-diagonal part of $\boldsymbol{X}$. The optimal $\tilde{\boldsymbol{B}}$ is then

$$\tilde{\boldsymbol{B}} = -\boldsymbol{D}_w^{1/2}\big(\boldsymbol{L} - \boldsymbol{D}(\boldsymbol{L})\big)\boldsymbol{L}^{-1}, \tag{8.7}$$

and the resulting minimum MSE of subchannel i is

$$[\boldsymbol{R}_{\mathrm{MSE}}\boldsymbol{D}_w]_{i,i} = [\boldsymbol{D}_w^{1/2}\boldsymbol{D}(\boldsymbol{L})]_{i,i}^2 = [\boldsymbol{D}_w]_{i,i}[\boldsymbol{L}]_{i,i}^2. \tag{8.8}$$

Note here that the impact the precoder $\boldsymbol{F}$ has on the weighted MSE is given entirely via the squared diagonal elements of the Cholesky decomposition, $\boldsymbol{L}$.

8.3 Optimal precoder: Left and right unitary matrices

The remaining filter to optimize is the precoder, $\boldsymbol{F}$. Consider the SVD of the precoder matrix

$$\boldsymbol{F} = \boldsymbol{U}_{\boldsymbol{F}}\boldsymbol{\Sigma}_{\boldsymbol{F}}\boldsymbol{V}_{\boldsymbol{F}}^*, \tag{8.9}$$

where $\boldsymbol{U}_{\boldsymbol{F}}$ and $\boldsymbol{V}_{\boldsymbol{F}}$ are unitary matrices, and $\boldsymbol{\Sigma}_{\boldsymbol{F}}$ is a non-negative diagonal matrix but with a non-specified ordering of the diagonal. For arbitrary monotonic increasing objective functions of the MSEs, the optimal left unitary matrix has been shown to match the matrix of eigenvectors of $\boldsymbol{H}^*\boldsymbol{H} = \boldsymbol{V}_{\boldsymbol{H}}\boldsymbol{\Lambda}_{\boldsymbol{H}}^2\boldsymbol{V}_{\boldsymbol{H}}^*$ [PCL03], [PJ07, Theorem 4.3]:

$$\boldsymbol{U}_{\boldsymbol{F}} = \boldsymbol{V}_{\boldsymbol{H}}, \tag{8.10}$$

where it is assumed that the diagonal of $\boldsymbol{\Lambda_H}$ is decreasing, *and* that the order of the diagonal of $\boldsymbol{\Sigma_F}$ satisfies that $\boldsymbol{\Lambda_H}\boldsymbol{\Sigma_F}$ is decreasing. A proof of this statement is also available in Lemma 2.5.1. Using (8.10), the right unitary matrix, $\boldsymbol{V_F}$, can readily be obtained from the SVD of $\boldsymbol{L}$ as

$$\begin{aligned} \boldsymbol{F}^*\boldsymbol{H}^*\boldsymbol{H}\boldsymbol{F}+\boldsymbol{I} = \boldsymbol{V_F}(\boldsymbol{\Sigma_F}^2\boldsymbol{\Lambda_H}^2+\boldsymbol{I})\boldsymbol{V_F}^* \implies \\ \boldsymbol{L} = \boldsymbol{V_F}(\boldsymbol{\Sigma_F}^2\boldsymbol{\Lambda_H}^2+\boldsymbol{I})^{-1/2}\boldsymbol{U}^*, \end{aligned} \tag{8.11}$$

where $\boldsymbol{U}$ is a unitary matrix that makes $\boldsymbol{L}$ lower triangular. Note here that the diagonal matrix with the singular values $\boldsymbol{\Lambda_L} \triangleq (\boldsymbol{\Sigma_F}^2\boldsymbol{\Lambda_H}^2+\boldsymbol{I})^{-1/2}$ is in this case ordered with increasing diagonal entries. Finally, also assuming $\boldsymbol{L}$ is known, the singular values of $\boldsymbol{F}$ are obtained from $\boldsymbol{\Lambda_L}$ and $\boldsymbol{\Lambda_H}$ as

$$\boldsymbol{\Sigma_F}^2 = \boldsymbol{\Lambda_H}^{-2}(\boldsymbol{\Lambda_L}^{-2}-\boldsymbol{I}). \tag{8.12}$$

With the optimal $\boldsymbol{W}^*$, $\boldsymbol{B}$, and $\boldsymbol{U_F}$, the remaining optimization problem is

$$\begin{aligned} \underset{\boldsymbol{V_F},\boldsymbol{U},\boldsymbol{\sigma}}{\text{minimize}} \quad & \left\|\boldsymbol{D}_w|\boldsymbol{d}(\boldsymbol{L})|^2\right\|_p && \text{(8.13a)} \\ \text{subject to} \quad & \boldsymbol{L} = \boldsymbol{V_F}(\boldsymbol{D}(\boldsymbol{\sigma})\boldsymbol{\Lambda_H}^2+\boldsymbol{I})^{-1/2}\boldsymbol{U}^*, && \text{(8.13b)} \\ & [\boldsymbol{L}]_{i,j} = 0 \quad \forall\, i<j, && \text{(8.13c)} \\ & \boldsymbol{V_F},\boldsymbol{U} \in \mathcal{U}, && \text{(8.13d)} \\ & \boldsymbol{\sigma} \geq 0, && \text{(8.13e)} \\ & \mathbf{1}^{\mathrm{T}}\boldsymbol{\sigma} \leq P, && \text{(8.13f)} \end{aligned}$$

where we introduced the power vector $\boldsymbol{\sigma} = \boldsymbol{d}(\boldsymbol{\Sigma_F}^2)$, that represents the power assigned to each spatial channel, and where $\mathcal{U}$ is the set of all N-dimensional unitary matrices.

Now, optimizing the unitary matrices $\boldsymbol{V_F}$, $\boldsymbol{U}$, directly is very difficult. From the expressions of the optimal DF filters we see that the filters either implicitly or explicitly depend on the lower triangular matrix $\boldsymbol{L}$. By optimizing $\boldsymbol{L}$ instead of $\boldsymbol{V_F}$, $\boldsymbol{U}$, we can avoid the unitary constraints. In order to do this we need a way to specify the singular values of $\boldsymbol{L}$ as an optimization constraint. Fortunately, this is possible since the diagonal elements of a triangular matrix are equal to its eigenvalues, and it is known that the absolute values of the eigenvalues are always multiplicatively majorized by the singular values (cf. [HJ91]). Interestingly, this necessary condition is also a sufficient condition on the triangular matrix $\boldsymbol{L}$: For a given power load, $\boldsymbol{\Sigma_F}^2$, and a specified vector $\boldsymbol{d}(\boldsymbol{L})$, one can uniquely determine (using generalized triangular decomposition (GTD) [JHL08]) the lower triangular matrix $\boldsymbol{L}$ if and only if

$$|\boldsymbol{d}(\boldsymbol{L})|^{-2} \preceq_{\times} \boldsymbol{d}(\boldsymbol{\Sigma_F}^2\boldsymbol{\Lambda_H}^2+\boldsymbol{I}), \tag{8.14}$$

where $\preceq_{\times}$ denotes multiplicative majorization [MO79, PJ07]. The proof of this statement was given in [JHL08], and for completeness, in Appendix 8.A a short

introduction to GTD is provided. With this necessary and sufficient condition on the diagonal elements of a triangular matrix we can replace the constraints (8.13b), (8.13c), and (8.13d) with the majorization constraint (8.14). Matrix notation becomes tedious (and unnecessary) to work with at this point, instead define the vectors $\boldsymbol{w} = \log \boldsymbol{d}(\boldsymbol{D}_w)$, $\boldsymbol{\xi} = \log |\boldsymbol{d}(\boldsymbol{L})|^{-2}$, and use (8.14) to pose the equivalent optimization problem in vector notation as

$$\begin{aligned} \underset{\boldsymbol{\xi},\boldsymbol{\sigma}}{\text{minimize}} \quad & \| \exp(\boldsymbol{w} - \boldsymbol{\xi}) \|_p && \text{(8.15a)} \\ \text{subject to} \quad & \boldsymbol{\xi} \preceq \log(\boldsymbol{\Lambda}_{\boldsymbol{H}}^2 \boldsymbol{\sigma} + \mathbf{1}), && \text{(8.15b)} \\ & \boldsymbol{\sigma} \geq 0, && \text{(8.15c)} \\ & \mathbf{1}^{\mathrm{T}} \boldsymbol{\sigma} \leq P, && \text{(8.15d)} \end{aligned}$$

where $\preceq$ denotes additive majorization. The vector $\boldsymbol{w}$ relates to the MSE weights of the problem and we call it the log-weights vector. The vector $\boldsymbol{\xi}$ has the interpretation of data rate (see (8.15b)) and we name it rate vector[3].

8.4 Optimal precoder: Power allocation

The optimization problem (8.15) is somewhat difficult due to the majorization constraint. Fortunately a simpler convex problem can be considered instead, as will be shown in this section. Because the p-norm is symmetric and because a majorization inequality is invariant to permutations of the vector elements, we can without loss of generality assume the subchannels are ordered such that both $\boldsymbol{w}$ and $\boldsymbol{\lambda} \triangleq \boldsymbol{d}(\boldsymbol{\Lambda}_{\boldsymbol{H}}^2)$ are decreasing.

Now that we have fixed the ordering of $\boldsymbol{w}$ and $\boldsymbol{\lambda}$, we may consider the following optimization problem

$$\begin{aligned} \underset{\boldsymbol{\xi},\boldsymbol{\sigma}}{\text{minimize}} \quad & \| \exp(\boldsymbol{w} - \boldsymbol{\xi}) \|_p && \text{(8.16a)} \\ \text{subject to} \quad & \sum_{j=1}^{i} \Big(\xi_j - \log(1 + \sigma_j \lambda_j) \Big) \leq 0 \quad \forall\, i, && \text{(8.16b)} \\ & \boldsymbol{\sigma} \geq 0, && \text{(8.16c)} \\ & \mathbf{1}^{\mathrm{T}} \boldsymbol{\sigma} \leq P. && \text{(8.16d)} \end{aligned}$$

Note that Problem (8.15) differs from (8.16), in that the majorization constraint has been replaced with another (similar) constraint that does not include the monotonic rearrangement of the vector elements. Although the two problems seem to be different, the following theorem states that the both problems share the same optimal solution.

[3]Note also, that $\boldsymbol{\xi}$ is the logarithm of the inverse of the MSEs, which corresponds to the mutual information over AWGN channels using Gaussian codebooks.

Theorem 8.4.1 *Given that* $\boldsymbol{\lambda}$ *and* $\boldsymbol{w}$ *are decreasing, the optimum solutions of problems* (8.15) *and* (8.16) *coincide.*

Proof: First we show that if $\boldsymbol{\lambda}$ is decreasing, then the optimal power loading, $\boldsymbol{\sigma}$, for problems (8.15) and (8.16) must ensure that $\boldsymbol{\Lambda}_{\boldsymbol{H}}^2\boldsymbol{\sigma}$ is decreasing: Assume that $\boldsymbol{\Lambda}_{\boldsymbol{H}}^2$ has a strictly positive diagonal. Define $\boldsymbol{\alpha} \triangleq \boldsymbol{\Lambda}_{\boldsymbol{H}}^2\boldsymbol{\sigma}$. Let $\boldsymbol{\Pi}$ be an arbitrary permutation matrix. Define an alternative power allocation, $\tilde{\boldsymbol{\sigma}}$, that yields a permutation of $\boldsymbol{\alpha}$ as $\boldsymbol{\Lambda}_{\boldsymbol{H}}^2\tilde{\boldsymbol{\sigma}} \triangleq \boldsymbol{\Pi}\boldsymbol{\alpha}$. Now, the total power consumption for the alternative power allocation is $\mathbf{1}^{\mathrm{T}}\tilde{\boldsymbol{\sigma}} = \mathbf{1}^{\mathrm{T}}\boldsymbol{\Lambda}_{\boldsymbol{H}}^{-2}\boldsymbol{\Pi}\boldsymbol{\alpha}$. Because $\boldsymbol{\Lambda}_{\boldsymbol{H}}^{-2}\mathbf{1}$ is increasing by assumption, the permutation matrix that yields the minimum power consumption is the one that makes $\boldsymbol{\Pi}\boldsymbol{\alpha}$ decreasing. Consequently, if $\boldsymbol{\alpha}$ is not decreasing then it cannot be optimal, and for both problems the optimal solution yields a decreasing vector $\log(\boldsymbol{\Lambda}_{\boldsymbol{H}}^2\boldsymbol{\sigma} + \mathbf{1})$.

As a consequence, by forcing the vector $\boldsymbol{\Lambda}_{\boldsymbol{H}}^2\boldsymbol{\sigma}$ to be decreasing in both problems we do not change their corresponding optima. The two problems can therefore be reformulated with more strict constraints: The reformulated version of Problem (8.15) is

$$\underset{\boldsymbol{\xi},\boldsymbol{\sigma}}{\text{minimize}} \quad \|\exp(\boldsymbol{w}-\boldsymbol{\xi})\|_p \tag{8.17a}$$

$$\text{subject to} \quad \sum_{j=1}^{i}\xi_{[j]} - \log(1+\lambda_j\sigma_j) \le 0 \quad \forall\, i, \tag{8.17b}$$

$$\sum_{i=1}^{N}\xi_i = \sum_{i=1}^{N}\log(1+\lambda_i\sigma_i), \tag{8.17c}$$

$$\lambda_i\sigma_i \ge \lambda_{i+1}\sigma_{i+1} \quad \forall\, 1 \le i < N, \tag{8.17d}$$

$$\boldsymbol{\sigma} \ge 0, \tag{8.17e}$$

$$\mathbf{1}^{\mathrm{T}}\boldsymbol{\sigma} \le P, \tag{8.17f}$$

and the corresponding reformulation of (8.16) is

$$\underset{\boldsymbol{\xi},\boldsymbol{\sigma}}{\text{minimize}} \quad \|\exp(\boldsymbol{w}-\boldsymbol{\xi})\|_p \tag{8.18a}$$

$$\text{subject to} \quad \sum_{j=1}^{i}\xi_j - \log(1+\sigma_j\lambda_j) \le 0 \quad \forall\, i, \tag{8.18b}$$

$$\sum_{j=1}^{N}\xi_j - \log(1+\sigma_j\lambda_j) = 0, \tag{8.18c}$$

$$\lambda_i\sigma_i \ge \lambda_{i+1}\sigma_{i+1} \quad \forall\, 1 \le i < N, \tag{8.18d}$$

$$\boldsymbol{\sigma} \ge 0, \tag{8.18e}$$

$$\mathbf{1}^{\mathrm{T}}\boldsymbol{\sigma} \le P. \tag{8.18f}$$

Note that the equality constraint (8.18c) is a consequence of the objective (8.18a) being decreasing with respect to ξ_N, so that the optimal ξ_N must achieve the upper bound defined by the only inequality containing ξ_N.

From the definition of monotonic rearrangements we have

$$\sum_{j=1}^{i} \xi_j \leq \sum_{j=1}^{i} \xi_{[j]}, \tag{8.19}$$

and we see that Problem (8.18) is a relaxation of (8.17). Furthermore, since the function $\|\exp(-\boldsymbol{z})\|_p$ is Schur-convex with respect to $\boldsymbol{z}$, and because $\boldsymbol{w}$ and $\log(\boldsymbol{\Lambda}_{\boldsymbol{H}}^2\boldsymbol{\sigma}+\mathbf{1})$ are decreasing, it can be shown, due to the constraints (8.18b), that the optimal $\boldsymbol{\xi}$ in (8.18) will be decreasing. The proof is given later in Theorem 2, Chapter 10. This means that the optimum of the relaxed problem (8.18) is also a feasible point given the constraints in (8.17). Hence, the problems (8.17), (8.18), (8.15), and (8.16) have equivalent optimal solutions. □

Problem (8.16) is convex and can be solved numerically with relative ease using standard tools for convex optimization [BV04]. In Section 8.5, we present an algorithm that solves the problem exactly with only $O(N)$ complexity.

To summarize, Figure 8.1 shows the flow chart for calculating the optimal DF filters. The first step is to compute the power loading, $\boldsymbol{\sigma}$, and rate vector, $\boldsymbol{\xi}$, e.g., by using the algorithm in Section 8.5. The second step uses the generalized triangular decomposition to compute the Cholesky factor, $\boldsymbol{L}$. The third step computes both the precoder, $\boldsymbol{F}$, and the feedback filter, $\boldsymbol{B}$, from $\boldsymbol{L}$. Finally, the feed forward filter, $\boldsymbol{W}^*$, is computed from $\boldsymbol{F}$ and $\boldsymbol{B}$. Note that after the first step we can already evaluate the objective value. It is therefore less computationally demanding to evaluate the performance than it is to compute the optimal filters for a particular bit loading. This fact allows us to reduce the complexity when an exhaustive search for the jointly optimal bit loading is performed (as was proposed in Section 7.4).

8.5 Algorithm that solves Problem (8.16)

Algorithm 1 presented below solves Problem (8.16) exactly with only $O(N)$ complexity. The proof of this statement is available in Appendix 8.B. The algorithm has some similarities with an algorithm presented in [JHL06] that solves the quality of service (QoS) constrained MSE optimization problem.

The input to the algorithm is the decreasing vectors $\boldsymbol{w}$ and $\boldsymbol{\lambda}$ of length N, and the p parameter. In the algorithm, the arrow '$\leftarrow$' denotes assignment of a variable, '&' and '|' denotes the logical operators AND and OR respectively. When the algorithm terminates, the result set consists of the following variables: q, α, $I_0, ..., I_q$, and $C_0, ..., C_{q-1}$. The sequence $I_0, ..., I_q$ is an ordered subset of the indices $0, ..., N$ and it defines the indices for which the constraints (8.16b) are satisfied with equality. Within each interval $I_i + 1, ..., I_{i+1}$ the optimal power allocation follows a waterfilling-like solution where the water levels are given by $e^{C_i - \alpha}$. From the result

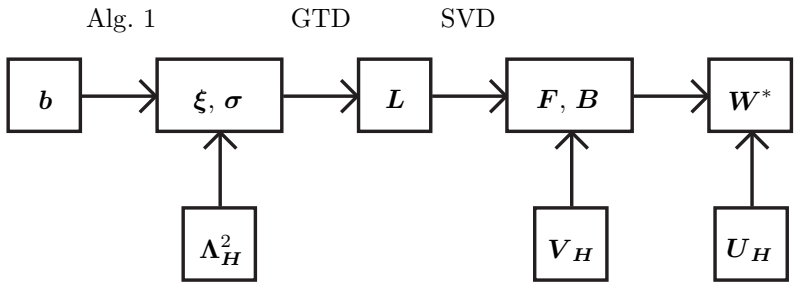

Figure 8.1: Flow chart for the calculation of the optimal DF filters given a bit loading vector $\boldsymbol{b}$. Alg. 1 denotes the algorithm in Appendix 8.5, GTD denotes generalized triangular decomposition.

set we can calculate the optimal power assignments $\sigma_1, ..., \sigma_N$, and MSE exponents $\xi_1, ..., \xi_N$, as:

- For all $j = I_i + 1, ..., I_{i+1}$, and for all $i = 0, ..., q-1$, assign

$$\sigma_j = \left(e^{C_i - \alpha} - \lambda_j^{-1}\right)^+, \quad \xi_j = w_j - p^{-1}C_i - \alpha. \tag{8.20}$$

- For all $j = I_q + 1, ..., N$, assign $\sigma_j = 0, \xi_j = 0$.

In the algorithm two lines are tagged as 'Pos X' and 'Pos Y', these are merely comments that will be used as references in the proof related to the algorithm.

Algorithm 1 Power allocation algorithm

Allocate memory for the following vectors of length $N+1$:
$I[0], ..., I[N]$, $C[0], ..., C[N]$, $K[0], ..., K[N]$.
Initialization of variables:
$I[0] \leftarrow 0$, $q \leftarrow 0$, $n \leftarrow 0$, $k \leftarrow 0$,
$s \leftarrow 0$, $A \leftarrow$ true, $B \leftarrow$ true.
For all $i = 1, ..., N$:
$\breve{w}_i \leftarrow \sum_{j=1}^{i} w_j$, $\breve{g}_i \leftarrow \sum_{j=1}^{i} \lambda_j^{-1}$, $\breve{h}_i \leftarrow \sum_{j=1}^{i} \log \lambda_j^{-1}$,
$\breve{w}_0 = \breve{h}_0 = \breve{g}_0 = 0$.
while $A \,|\, B$ **do**
  Given $I[q]$ and n, calculate the highest $k \in \left\{ I[q]+1, ..., n \right\}$ such that $\gamma \geq \log \lambda_k^{-1} + (1+\beta)\alpha$, where

$$\gamma = \frac{\breve{w}_n - \breve{w}_{I[q]} + \breve{h}_k - \breve{h}_{I[q]}}{k + np^{-1} - (1+p^{-1})I[q]},$$

$$\beta = \frac{(n-k)}{k + np^{-1} - (1+p^{-1})I[q]},$$

  and finally α that is evaluated by solving the non-linear equation

$$(P + \breve{g}_k)e^{\alpha} = s + (k - I[q])\, e^{\gamma - \alpha\beta}.$$

  $K[q] \leftarrow k - I[q]$
  $C[q] \leftarrow \gamma - \beta\alpha$
  % Pos Y:
  if $(q > 0)\,\&\,(C[q-1] \leq C[q])$ **then**
    $q \leftarrow q - 1$
    $s \leftarrow s - K[q]\, e^{C[q]}$
  else
    if $n < N$ **then**
      $A \leftarrow (k = n)\,\&\left(w_{n+1} - \alpha > p^{-1}(\log \lambda_{n+1}^{-1} + \alpha)\right)$
      $B \leftarrow (p^{-1}C[q] < w_{n+1} - \alpha)$
    else
      $A \leftarrow$ false, $B \leftarrow$ false
    end if
    % Pos X:
    if $A \,|\, \boldsymbol{n}ot(B)$ **then**
      $s \leftarrow s + K[q]\, e^{C[q]}$
      $q \leftarrow q + 1$
      $I[q] \leftarrow n$
    end if
    $n \leftarrow n + 1$
  end if
end while

Appendix 8.A Generalized triangular decomposition

The generalized triangular decomposition (GTD) [JHL08] is a matrix decomposition that can be seen as a generalization of many well-known decompositions including; the singular value decomposition [Bel73, Jor74], the QR factorization [Giv58, Hou58], and the geometric mean decomposition [KS00, JHL05]. A rank K matrix $\boldsymbol{X} \in \mathbb{C}^{M\times N}$ is decomposed as

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{R}\boldsymbol{V}^*, \tag{8.21}$$

where $\boldsymbol{R} \in \mathbb{C}^{K\times K}$ is upper triangular, and $\boldsymbol{U}$, $\boldsymbol{V}$ are matrices with orthonormal columns. From [JHL08] we have the following theorem:

Theorem 8.A.1 *Let $\boldsymbol{X} \in \mathbb{C}^{M\times N}$ have rank $\boldsymbol{K}$ with singular values $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_K \geq 0$. There exists an upper triangular matrix $\boldsymbol{R} \in \mathbb{C}^{K\times K}$ with diagonal prescribed by $\boldsymbol{r}$, and matrices $\boldsymbol{U}$, $\boldsymbol{V}$ with orthonormal columns such that*

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{R}\boldsymbol{V}^*, \tag{8.22}$$

if and only if $|\boldsymbol{r}| \preceq_\times \boldsymbol{\sigma}$.

Along with this theorem, an algorithm that produces $\boldsymbol{R}$ given $\boldsymbol{r}$ and $\boldsymbol{\sigma}$ is presented. Below we give a brief overview on the main steps of the algorithm.

First, we note that by using the SVD of $\boldsymbol{X} = \boldsymbol{U}_{\boldsymbol{X}}\boldsymbol{\Sigma}\boldsymbol{V}_{\boldsymbol{X}}^*$, an equivalent problem is to find the unitary matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ such that

$$\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \boldsymbol{R}, \tag{8.23}$$

where the diagonal of the upper triangular matrix $\boldsymbol{R}$ is prescribed. Now, consider the simplified case with matrices of dimension 2×2, and denote the diagonal elements of $\boldsymbol{\Sigma}$ as σ_1 and σ_2. By using the following parametrization

$$c = \sqrt{\frac{|r_1|^2 - |\sigma_2|^2}{|\sigma_1|^2 - |\sigma_2|^2}}, \qquad s = \sqrt{1-c^2}, \tag{8.24}$$

we construct a unitary matrix, $\boldsymbol{V}$, as

$$\boldsymbol{V} = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}. \tag{8.25}$$

Then, we perform the QR decomposition of $\boldsymbol{\Sigma}\boldsymbol{V}^* = \boldsymbol{U}^*\boldsymbol{R}$, where $\boldsymbol{U}$ is a unitary matrix, to obtain the GTD as

$$\underbrace{\frac{1}{r_1^*}\begin{bmatrix} c\sigma_1^* & s\sigma_2^* \\ -s\sigma_2 & c\sigma_1 \end{bmatrix}}_{\boldsymbol{U}} \underbrace{\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}}_{\boldsymbol{\Sigma}} \underbrace{\begin{bmatrix} c & -s \\ s & c \end{bmatrix}}_{\boldsymbol{V}^*} = \underbrace{\begin{bmatrix} r_1 & \times \\ 0 & r_2 \end{bmatrix}}_{\boldsymbol{R}}. \tag{8.26}$$

The same procedure can be applied to matrices of higher dimension. By multiplying the given diagonal matrix $\boldsymbol{\Sigma}$ with unitary matrices that linearly combine a pair of

$$\left[\begin{array}{cc|cccc} r_1 & \times & \times & \times & \times & \times \\ 0 & r_2 & \times & \times & \times & \times \\ \hline 0 & 0 & \sigma_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_6 \end{array}\right] \qquad \left[\begin{array}{ccc|ccc} r_1 & \times & \times & \times & \times & \times \\ 0 & r_2 & \times & \times & \times & \times \\ 0 & 0 & r_3 & \times & \times & \times \\ \hline 0 & 0 & 0 & \sigma_4' & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_5' & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_6' \end{array}\right]$$

Figure 8.2: One iteration in the GTD algorithm. The left-hand side shows the state of the $\boldsymbol{R}$ matrix before multiplying with a unitary matrix pair that enforces the third diagonal element to the prescribed value r_3. The result is shown in the right-hand side. Note how the lower right submatrix remains diagonal but with modified diagonal entries.

diagonal elements i, j, we will get a new triangular matrix $\boldsymbol{R}$. The diagonal elements of $\boldsymbol{R}$ remains unchanged except on the positions i, and j, for which the diagonal elements are r_1 and r_2. By repeating the procedure for elements $i = 1, ..., N-1$ and $j > i$ we can transform the diagonal from $\boldsymbol{\sigma}$ to $\boldsymbol{r}$ one element at a time. Figure 8.2 shows one iteration in the GTD algorithm.

Perhaps the most fascinating aspect of the GTD algorithm is that, by rotating a matrix $\boldsymbol{X}$ with a unitary matrix $\boldsymbol{W}$, we can change the eigenvalues of $\boldsymbol{X}$ arbitrarily (to $\boldsymbol{\lambda}$) provided the majorization constraint $\boldsymbol{\lambda} \preceq_\times \boldsymbol{\sigma}$ is satisfied: Say we want to change the eigenvalues to $\boldsymbol{\lambda}$. First compute the GTD of $\boldsymbol{X}$ such that the diagonal elements of $\boldsymbol{R}$ is $\boldsymbol{\lambda}$. We have

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{R}\boldsymbol{V}^*. \tag{8.27}$$

Then define the unitary matrix $\boldsymbol{W} = \boldsymbol{V}\boldsymbol{U}^*$ to obtain

$$\boldsymbol{X}\boldsymbol{W} = \boldsymbol{U}\boldsymbol{R}\boldsymbol{U}^* = \boldsymbol{U}\boldsymbol{R}\boldsymbol{U}^{-1}. \tag{8.28}$$

The diagonal elements of a triangular matrix $\boldsymbol{R}$ are equivalent to its eigenvalues, and since $\boldsymbol{U}\boldsymbol{R}\boldsymbol{U}^{-1}$ and $\boldsymbol{R}$ have identical eigenvalues, the eigenvalues of $\boldsymbol{X}\boldsymbol{W}$ are $\boldsymbol{\lambda}$.

Appendix 8.B Proof of Algorithm 1

In order to avoid numerical complications with the infinity norm, the norm coefficient may be redefined as

$$\phi = p^{-1} \in [0, 1]. \tag{8.29}$$

Consider the optimization problem in (8.16). Since the problem is convex, the KKT optimality conditions [BV04] are necessary and sufficient conditions for optimality.

Using $\epsilon_1, ..., \epsilon_N$ and δ as duality variables the KKT conditions can be posed as[4]

$$\xi_i = w_i - \phi \log\Big(\phi \sum_{j=i}^{N} \epsilon_j\Big), \tag{8.30a}$$

$$\sigma_i = \Big(\delta^{-1} \sum_{j=i}^{N} \epsilon_j - \lambda_i^{-1}\Big)^+, \tag{8.30b}$$

$$\epsilon_i \geq 0, \quad \forall\, i, \tag{8.30c}$$

$$\epsilon_i \Big(\sum_{j=1}^{i} \xi_j - \log(1 + \sigma_j \lambda_j) \Big) = 0, \quad \forall\, i, \tag{8.30d}$$

$$\sum_{j=1}^{i} \xi_j - \log(1 + \sigma_j \lambda_j) \leq 0, \quad \forall\, i, \tag{8.30e}$$

$$\mathbf{1}^{\mathrm{T}} \boldsymbol{\sigma} = P. \tag{8.30f}$$

The core when solving these conditions is to identify the set of indices where $\epsilon_i > 0$. Once these are known, the remaining conditions follows straightforwardly, as will be shown. Denote these indices $I_1, ..., I_Q$, and define also $I_0 = 0$. As a first step we will refine these KKT conditions to a set of conditions that are more easy to handle, then we will show (inductively) that the result set of the algorithm will satisfy the conditions.

The summation $\sum_{j=i}^{N} \epsilon_j$ that appears in (8.30a) and (8.30b) remains constant for indices i in intervals, defined by $I_1, ..., I_Q$ as

$$i \in \mathcal{I}_r \triangleq \{1 + I_r, ..., I_{r+1}\}, \tag{8.31}$$

where r is the interval index and ranges from 0 to $Q - 1$. To see this, note that

$$\epsilon_i = 0 \quad \forall \ \in \mathcal{I}_r \setminus \{I_{r+1}\}. \tag{8.32}$$

Now, introduce the following interval coefficients

$$C_r = \log\left(\phi \sum_{j=1+I_r}^{N} \epsilon_j\right) - \frac{\alpha}{\phi}, \quad \forall\, r = 0, ..., Q - 1, \tag{8.33}$$

where the variable α is defined as

$$\alpha = \frac{\phi}{\phi + 1} \log(\delta\phi). \tag{8.34}$$

[4]For simplicity we have eliminated the possibility of $\delta = 0$. It is easy to verify that this is not a solution to the KKT conditions; $\delta = 0 \Rightarrow \epsilon_i = 0\, \forall\, i$, which leads to infinite power, which in turn contradicts the fact that $\delta = 0$.

With these definitions in place, we can make substitutions into the conditions (8.30) for ξ_i and σ_i as

$$\left.\begin{aligned}\xi_i &= w_i - \alpha - \phi\, C_r \\ \sigma_i &= \left(e^{C_r-\alpha} - \lambda_i^{-1}\right)^+\end{aligned}\right\} \quad \forall\, i \in \mathcal{I}_r,\ \forall\, r. \tag{8.35}$$

Furthermore, the condition $\epsilon_i \geq 0$ for all i combined with the definition of the intervals can be replaced by the condition that C_r is strictly decreasing with r. Consider condition (8.30d), because we know that ϵ_i is zero for all i except for $i \in \{I_1, I_2, ..., I_Q\}$, for which ϵ_i is strictly positive, we must have

$$\sum_{j\in\mathcal{I}_r} \xi_j - \log(1+\sigma_j\lambda_j) = 0, \quad \forall\, r = 0, ..., Q-1. \tag{8.36}$$

From the above substitutions we can simplify the KKT optimality conditions to

$$\xi_i = w_i - \alpha - \phi\, C_r, \quad \forall\, i \in \mathcal{I}_r, \quad \forall\, r, \tag{8.37a}$$

$$\sigma_i = \left(e^{C_r-\alpha} - \lambda_i^{-1}\right)^+, \quad \forall\, i \in \mathcal{I}_r, \quad \forall\, r = 0, ..., Q-1, \tag{8.37b}$$

$$\sum_{j\in\mathcal{I}_r} \xi_j - \log(1+\sigma_j\lambda_j) = 0, \quad \forall\, r, \tag{8.37c}$$

$$\mathbf{1}^{\mathrm{T}}\boldsymbol{\sigma} = P. \tag{8.37d}$$

$$C_0 > C_1 > ... > C_{Q-1}, \tag{8.37e}$$

$$\sum_{j=1}^{i} \xi_j - \log(1+\sigma_j\lambda_j) \leq 0, \quad \forall\, i, \tag{8.37f}$$

If we ignore the inequalities in (8.37), we have in total $3Q+1$ equations and $3Q+1$ unknowns, so given a set of indices $I_1, ..., I_Q$ we also have a unique solution. The problem is to find the set of indices $I_1, ..., I_Q$ that gives a solution that also satisfies the inequalities. We will find this set inductively by assuming the optimal set $I_1, ..., I_q, I_{q+1}$ has been obtained for a system of size n, then add one extra element and see how this changes the optimal index set.

The algorithm steps through the indices $n = 1, ..., N$ one by one. In each step the optimality conditions will be ensured for the indices $1, ..., n$. At each specific step, n, we have $q+1$ intervals defined by the index sequence $I_1, ..., I_q, n$. Note that by assuming we have optimality, both C_r and λ_i are decreasing, and hence the sequence $e^{C_r-\alpha} - \lambda_i^{-1}$ must be decreasing too. This implies that there is exactly one index k that is the highest index with a non-zero power allocation σ_i. This index is denoted the water-level index, and it must satisfy

$$\log \lambda_{k+1}^{-1} > C_r - \alpha \geq \log \lambda_k^{-1}. \tag{8.38}$$

If $C_q - \alpha \geq \log \lambda_n^{-1}$ then $k = n$. In the following we assume $k > I_q$ because, as will be shown later, when $k \leq I_q$, our problem is solved since the remaining set of indices $I_{q+1}, ..., I_Q$ can be obtained instantly.

For notational simplicity in the lemmas and theorems below, we introduce the following definitions and relations

$$h_i = \log \lambda_i^{-1}, \tag{8.39}$$

$$\bar{k}_j = \min(j, k), \tag{8.40}$$

$$K_r = \min(k, I_{r+1}) - I_r, \tag{8.41}$$

$$f(j, k, I_r) = \sum_{i=I_r+1}^{j} w_i + \sum_{i=I_r+1}^{\bar{k}_j} h_i, \tag{8.42}$$

$$g(j, k, I_r) = (\bar{k}_j - I_r) + \phi(j - I_r), \tag{8.43}$$

where we use the following (integral-style) convention for reversing the limits of summation

$$\sum_{i=a+1}^{b} x_i = \sum_{i=1}^{b} x_i - \sum_{i=1}^{a} x_i = -\sum_{i=b+1}^{a} x_i. \tag{8.44}$$

Applying these definitions to the equations in (8.37) we can isolate $C_0, ..., C_q$ from the ξ_i's and σ_i's as

$$\begin{aligned} C_r &= \frac{f(I_{r+1}, k, I_r)}{g(I_{r+1}, k, I_r)} \quad \forall\, r < q, \\ C_q &= \frac{f(n, k, I_q) - (n-k)\alpha}{g(n, k, I_q)}. \end{aligned} \tag{8.45}$$

Again using the new notation, KKT condition (8.30f) can be posed as

$$P + \sum_{i=1}^{k} e^{h_i} = \sum_{r=0}^{q} K_r e^{C_r - \alpha}, \tag{8.46}$$

which will be referred to as the power equation. So, given the index set $I_1, ..., I_q, n$, and a water-level index k, we can uniquely compute $C_0, ..., C_q$ and α from equations (8.45) and (8.46). In order for this solution to satisfy the KKT optimality conditions the inequalities in (8.37) as well as the inequalities relating to k needs to be fulfilled:

- Firstly, C_r has to be decreasing

$$C_0 > C_2 > ... > C_q. \tag{8.47a}$$

- Secondly, the water-level, k, index needs to satisfy

$$\begin{aligned} k < n: &\quad h_{k+1} > C_q - \alpha \geq h_k \\ k = n: &\quad C_q - \alpha \geq h_n \end{aligned} . \tag{8.47b}$$

- Thirdly, for all $j = I_q + 1, ..., n$, we need

$$f(j, k, I_q) \leq g(j, k, I_q)C_q + (j - \bar{k}_j)\alpha. \tag{8.47c}$$

- Finally, for all $j \in \mathcal{I}_r$ and all $r < q$

$$f(j, k, I_r) \leq g(j, k, I_r)C_r. \tag{8.47d}$$

The following lemmas and theorems provide a proof that Algorithm 1 produces a sequence $I_1, ..., I_Q$, with the corresponding $C_0, ..., C_{Q-1}$ and α that satisfies the optimality conditions (8.47). Most lemmas and theorems will compare the state of the variables on two time instances (positions) in the algorithm. The variables relating to the first position (in chronological order) will be denoted as in the algorithm description, i.e. n, q, C_q, γ, etc., while the variables relating to the later position will be denoted as primed, i.e. n', q', $C'_{q'}$, γ', etc. Note also that mixtures will occur: For instance C'_q means the value of variable C_q in the second chronological position, but with the index q with the value as it is on the first position.

Lemma 8.B.1 *If A is true at the line 'Pos X' in Algorithm 1 (in the following we will simply refer to Pos X and Pos Y for the corresponding lines in the algorithm), then the state variables will satisfy $k = n$ the following stop at position Pos Y, i.e. using the prime notation introduced above we have $k' = n'$.*

Proof: If A is true at Pos X, then $I_{q'} + 1 = n'$ and consequently $k' = n'$. However, we need to check that this k' satisfies $\gamma' \geq h_{k'} + \alpha'$ (otherwise k' is not defined in the algorithm): Using the condition for A to be true we have

$$\gamma' = \frac{w_{n+1} + h_{n+1}}{1 + \phi} > h_{n+1} + \alpha. \tag{8.48}$$

The power equations at Pos X and Pos Y respectively are

$$(P + \breve{g}_n)e^{\alpha} = s', \tag{8.49}$$

$$(P + \breve{g}_n + e^{h_{n+1}})e^{\alpha'} = s' + e^{\gamma'}. \tag{8.50}$$

Combining these equations with the γ' inequality we get

$$(P + \breve{g}_n + e^{h_{n+1}})e^{\alpha'} > (P + \breve{g}_n + e^{h_{n+1}})e^{\alpha}, \tag{8.51}$$

and $\alpha' > \alpha$. So we have shown that $\gamma' > h_{n+1} + \alpha'$ and thus k' is well defined and equal to n' at the following Pos Y. □

Lemma 8.B.2 *Between two stops at Pos X where $k < n$ and $k' \neq k$, we have the following inequalities*

$$f(j,k',I_r) \leq f(j,k,I_r) + (\bar{k}'_j - \bar{k}_j)\big(C'_{q'} - \alpha'\big), \tag{8.52a}$$
$$f(j,k',I_r) \geq f(j,k,I_r) + (\bar{k}'_j - \bar{k}_j)\big(C_q - \alpha\big). \tag{8.52b}$$

Proof: From definitions (8.42) and (8.47b) we have, assuming $k' > k$:

$$f(j,k',I_r) - f(j,k,I_r) = \sum_{i=\bar{k}_j+1}^{\bar{k}'_j} h_i \leq (\bar{k}'_j - \bar{k}_j)\big(C'_{q'} - \alpha'\big), \tag{8.53}$$

and similarly

$$f(j,k',I_r) - f(j,k,I_r) = \sum_{i=\bar{k}_j+1}^{\bar{k}'_j} h_i \geq (\bar{k}'_j - \bar{k}_j)\big(C_q - \alpha\big). \tag{8.54}$$

Assuming $k' < k$, we first note that $k' < n < n'$ and we have

$$f(j,k',I_r) - f(j,k,I_r) = -\sum_{i=\bar{k}'_j+1}^{\bar{k}_j} h_i \leq (\bar{k}'_j - \bar{k}_j)\big(C'_{q'} - \alpha'\big), \tag{8.55}$$

and similarly

$$f(j,k',I_r) - f(j,k,I_r) = -\sum_{i=\bar{k}'_j+1}^{\bar{k}_j} h_i \geq (\bar{k}'_j - \bar{k}_j)\big(C_q - \alpha\big). \tag{8.56}$$

□

Lemma 8.B.3 *At Pos Y, if $C_{q-1} \leq C_q$ we perform a merge in the following if-statement. The result (the consecutive stop at Pos Y) will then satisfy $C_{q-1} \leq C'_{q-1}$.*

Proof: Whenever $C_{q-1} \leq C_q$, a merge will take place between intervals $\{I_{q-1} + 1, ..., I_q\}$ and $\{I_q + 1, ..., n\}$. Similar to the proof in Lamma 8.B.2 we have

$$\begin{aligned} f(n,k',I_{q-1}) &\geq f(I_q,k,I_{q-1}) + f(n,k,I_q) + (k'-k)\big(C_q - \alpha\big) \\ &\geq g(n,k',I_{q-1})C_{q-1} + (n-k')\alpha, \end{aligned} \tag{8.57}$$

where the last inequality use (8.45). So then, if $\alpha \geq \alpha'$ we have $C'_{q-1} \geq C_{q-1}$. If on the other hand $\alpha' \geq \alpha$, then the power equation gives

$$\begin{aligned} \Big(P + \sum_{i=1}^{k'} e^{h_i}\Big) e^{\alpha'} &= K'_{q-1} e^{C'_{q-1}} + \sum_{r=1}^{q-2} K_r e^{C_r} \\ \Big(P + \sum_{i=1}^{k'} e^{h_i}\Big) e^{\alpha} &\leq K'_{q-1} e^{C'_{q-1}} + \sum_{r=1}^{q-2} K_r e^{C_r} \end{aligned} \tag{8.58}$$

At the same time, using the fact that $C_q \geq C_{q-1}$ along with (8.47b),

$$\begin{aligned} \Big(P + \sum_{i=1}^{k} e^{h_i}\Big) e^{\alpha} &= K_q e^{C_q} + K_{q-1} e^{C_{q-1}} + \sum_{r=1}^{q-2} K_r e^{C_r} \\ \Big(P + \sum_{i=1}^{k'} e^{h_i}\Big) e^{\alpha} &\geq K'_{q-1} e^{C_{q-1}} + \sum_{r=1}^{q-2} K_r e^{C_r} \end{aligned} \tag{8.59}$$

and consequently $C'_{q-1} \geq C_{q-1}$. □

Lemma 8.B.4 *On two consecutive stops at Pos X, if* (8.47a) *is satisfied at the first stop and if $q' < q$, then $C'_{q'} \geq C_{q'}$ and $\alpha' \geq \alpha$.*

Proof: Because we are at Pos X, we have $C_{q'-1} = C'_{q'-1} > C'_{q'}$. Since $q' < q$, at least one merge has occurred, from Lemma 8.B.3 we have $C'_{q'} \geq C_{q'}$. Since the sequence is valid initially we must have $C_{q'-1} > C'_{q'} \geq C_{q'} > ... > C_q$. Apply this result to the power equations

$$\begin{aligned} \Big(P + \sum_{i=1}^{k} e^{h_i}\Big) e^{\alpha'} &\geq \sum_{r=1}^{q'-1} K_r e^{C_r} + \Big(\sum_{r=q'}^{q} K_r\Big) e^{C'_{q'}}, \\ \Big(P + \sum_{i=1}^{k} e^{h_i}\Big) e^{\alpha} &\leq \sum_{r=1}^{q'-1} K_r e^{C_r} + \Big(\sum_{r=q'}^{q} K_r\Big) e^{C'_{q'}}, \end{aligned} \tag{8.60}$$

hence $\alpha' \geq \alpha$. □

Lemma 8.B.5 *On two consecutive stops at Pos X, if $q' = q$ then*

$$C'_q \geq C_q \iff \alpha' \geq \alpha. \tag{8.61}$$

Proof: Consider the power equations (8.46) and (8.38)

$$\begin{aligned}
\Big(P+\sum_{i=1}^{k} e^{h_i}\Big)e^{\alpha} &= \sum_{r=1}^{q-1} K_r e^{C_r} + K_q e^{C_q},\\
\Big(P+\sum_{i=1}^{k} e^{h_i}\Big)e^{\alpha'} &\geq \sum_{r=1}^{q-1} K_r e^{C_r} + K_q e^{C'_q},\\
\Big(P+\sum_{i=1}^{k'} e^{h_i}\Big)e^{\alpha'} &= \sum_{r=1}^{q-1} K_r e^{C_r} + K'_q e^{C'_q},\\
\Big(P+\sum_{i=1}^{k'} e^{h_i}\Big)e^{\alpha} &\geq \sum_{r=1}^{q-1} K_r e^{C_r} + K'_q e^{C_q}.
\end{aligned} \tag{8.62}$$

From these inequalities we can derive $C'_q \geq C_q \iff \alpha' \geq \alpha$. □

Lemma 8.B.6 *If A is true at Pos X and the state is optimal up to this point, then the state is optimal the next time the algorithm reaches Pos X.*

Proof: If $q' = q+1$, we have $C'_{q'-1} > C'_{q'}$ and we have $C'_r = C_r$ for all $r < q'$, thus condition (8.47a) is satisfied. From Lemma 8.B.1, we have $k' > k = I_{q'}$, conditions (8.47d) are satisfied. Finally, since $K'_{q'} = 1$, (8.47c) is satisfied.

If $q' \leq q$, then a number of merges have occurred. We have $C'_r = C_r$, for all $r < q'$ and $C'_{q'-1} > C'_{q'}$, and consequently condition (8.47a) is satisfied. Similarly, the conditions in (8.47d) are satisfied. Lemma 8.B.3 implies that $C'_{q'} > C_{q'}$, and because the state was optimal the first time at Pos X, we have

$$f(j,k,I_{q'}) \leq g(j,k,I_{q'})C'_{q'} + (j-\bar{k}_j)\alpha \tag{8.63}$$

Lemmas 8.B.4, 8.B.5, and 8.B.2 give

$$\begin{aligned}
f(j,k',I_{q'}) &\leq f(j,k,I_{q'}) + (\bar{k}'_j - \bar{k}_j)(C'_{q'} - \alpha')\\
&\leq g(j,k',I_{q'})C'_{q'} + (j-\bar{k}'_j)\alpha',
\end{aligned} \tag{8.64}$$

which satisfy condition (8.47c). □

Lemma 8.B.7 *Between two consecutive stops at Pos X, if A is false, B is true, and the state is optimal initially, then $\alpha' \geq \alpha$ and $C'_{q'} \geq C_{q'}$.*

Proof: Since A is false and B is true we have $q' \leq q$. For the case $q' < q$, we have $\alpha' \geq \alpha$ by Lemma 8.B.4. For the case $q' = q$, using $\phi C_q + \alpha < w_{n+1}$ and Lemma 8.B.2 we have

$$\begin{aligned} f(n+1,k',I_q) &> f(n,k,I_q) + (\phi + k' - k)C_q + \\ &\quad + (1 - k' + k)\alpha \implies \\ \frac{f(n+1,k',I_q) - (n+1-k')\alpha}{g(n+1,k',I_q)} &\geq \\ &\geq \frac{f(n,k,I_q) - (n-k)\alpha}{g(n,k,I_q)} = C_q. \end{aligned} \tag{8.65}$$

So if $\alpha' \leq \alpha$ then $C'_q \geq C_q$, which is a contradiction due to Lemma 8.B.5. Consequently $\alpha' \geq \alpha$ and $C'_{q'} \geq C_{q'}$ for all cases $q' \leq q$. □

Lemma 8.B.8 *If A is false and B is true at Pos X, and the state is optimal up to this point, then the state is optimal the next time the algorithm reaches Pos X.*

Proof: We will show that (8.47c) is valid. From Lemma 8.B.2 we have

$$f(j,k',I_{q'}) \leq f(j,k,I_{q'}) + (\bar{k}'_j - \bar{k}_j)\big(C'_{q'} - \alpha'\big) \tag{8.66}$$

and using the fact that (8.47c) is valid initially we have

$$f(j,k,I_{q'}) \leq g(j,k,I_{q'})C_{q'} + (j - \bar{k}_j)\alpha, \tag{8.67}$$

for all $j = I_{q'} + 1, ..., n$. By Lemma 8.B.7 we have $\alpha' \geq \alpha$ and $C'_{q'} \geq C_{q'}$ for all cases $q' \leq q$, hence (8.47c) is satisfied as

$$f(j,k',I_{q'}) \leq g(j,k',I_{q'})C'_{q'} + (j - \bar{k}'_j)\alpha', \tag{8.68}$$

for all $j = I_{q'} + 1, ..., n$. □

Lemma 8.B.9 *If* (8.47) *is satisfied up to $n \leq N$ and A as well as B are false at Pos X, then the algorithm has calculated the optimal index set and should terminate.*

Proof: If $\phi > 0$: For all $j = 1, ..., N - n$, construct the following elements

$$C_{q+j} = \frac{w_{n+j} - \alpha}{\phi}. \tag{8.69}$$

Note that (due to the fact that B is false and the decreasing sequence w_{n+1}) we have $C_q > C_{q+1} > ... > C_{q+N-n}$. Note also that since A is false, α will not be affected by these new elements. This means that we have found the optimal index set $I_1, ..., I_{q+N-n}$, and the algorithm can terminate.

If $\phi = 0$: We have reached the point where $w_{n+1} < \alpha$, it will not be possible to add any more elements. The optimal index set $I_1, ..., I_q$, has been obtained and the algorithm can terminate. □

Theorem 8.B.10 *Algorithm 1 produces the solution that satisfies the KKT conditions* (8.30).

Proof: Initially, $n = 1$ and

$$C_1 = \frac{w_1 + h_1}{1 + \phi}, \qquad \alpha = C_1 - \log(P + e^{h_1}). \tag{8.70}$$

Note that $C_1 \geq h_1 + \alpha$, so $k = n = 1$. Clearly since there is only one element C_1 with only one index $n = 1$, all optimality conditions (8.47a), (8.47c), and (8.47d), are satisfied at Pos X. By induction and using Lemmas 8.B.6 and 8.B.8 we then know that every time the algorithm passes by Pos X, the index set satisfies conditions (8.47a), (8.47c), and (8.47d). Finally, when eventually A is false and B is false, Lemma 8.B.9 tells us that the optimal index set has been obtained. □

Lemma 8.B.11 *Between two consecutive stops at Pos X, the water level index does not decrease $k' \geq k$.*

Proof: If A is true, by Lemma 8.B.1 we have $k' = k + 1$. If A is false and B is false, then the algorithm terminates and there is no next stop. If A is false and B is true, we have by Lemma 8.B.7 $q' \leq q$, $\alpha' \geq \alpha$ and $C'_{q'} \geq C_q$. Now assume $k' < k$, then $C'_{q'} - \alpha' < h_k < C_q - \alpha$. Consider the power equation

$$\begin{aligned} P + \sum_{i=1}^{k'} e^{h_i} &= K'_{q'} e^{C'_{q'} - \alpha'} + \sum_{r=1}^{q'-1} K_r e^{C_r - \alpha'} \\ &< K'_{q'} e^{C_q - \alpha} + \sum_{r=1}^{q'-1} K_r e^{C_r - \alpha}, \end{aligned} \tag{8.71}$$

$$P + \sum_{i=1}^{k} e^{h_i} < \Big(\sum_{r=q'}^{q} K_r \Big) e^{C_q - \alpha} + \sum_{r=1}^{q'-1} K_r e^{C_r - \alpha}.$$

Since (8.47a) is satisfied,

$$P + \sum_{i=1}^{k} e^{h_i} < \sum_{r=1}^{q} K_r e^{C_r - \alpha}, \tag{8.72}$$

which contradicts (8.46). Hence $k' \geq k$. □

Theorem 8.B.12 *Algorithm 1 has linear complexity.*

Proof: Between two consecutive stops at Pos X, n is increased by one until $n = N$, when the algorithm terminates. When the algorithm has finished there have been at most twice as many stops at Pos Y as there have been stops at Pos X since $q \geq 1$. By Lemma 8.B.11, k does not decrease between stops at Pos X, hence the search space for k does not grow with N. Consequently complexity of the algorithm scales linearly with N. □

Chapter 9

Optimal bit loading

This chapter switches focus to the bit loading problem, i.e. the problem of computing the optimal $\boldsymbol{b}$ in (7.15). In Chapter 8, it was shown that for any weighting vector $\boldsymbol{w}$, Problem (7.15) can be simplified to the form (8.15). Because the bit loading and the DF filters are coupled only through the cost function (there are no common constraints), the results from Chapter 8 can be incorporated into the original problem formulation (7.15) as

$$\begin{aligned}
\underset{\boldsymbol{\xi},\boldsymbol{\sigma},\boldsymbol{b}}{\text{minimize}} \quad & \| \exp(\boldsymbol{w} - \boldsymbol{\xi}) \|_p && \text{(9.1a)} \\
\text{subject to} \quad & \boldsymbol{\xi} \preceq \log(\boldsymbol{\Lambda}_{\boldsymbol{H}}^2 \boldsymbol{\sigma} + \mathbf{1}), && \text{(9.1b)} \\
& \boldsymbol{\sigma} \geq 0, && \text{(9.1c)} \\
& \mathbf{1}^{\mathrm{T}} \boldsymbol{\sigma} \leq P, && \text{(9.1d)} \\
& w_i = -\log d_{\min}^2(b_i) \quad \forall\, i, && \text{(9.1e)} \\
& b_i \in \mathcal{B} \quad \forall\, i, && \text{(9.1f)} \\
& \mathbf{1}^{\mathrm{T}} \boldsymbol{b} = R. && \text{(9.1g)}
\end{aligned}$$

In general, the set of available constellations, $\mathcal{B}$, is discrete. In particular, if QAM constellations are used the bit rates are restricted to positive, even integers. The set of feasible bit rates is then

$$\begin{aligned}
& b_i \in \{0, 2, 4, ...\} \,\forall\, i, && \text{(9.2a)} \\
& \mathbf{1}^{\mathrm{T}} \boldsymbol{b} = R. && \text{(9.2b)}
\end{aligned}$$

Clearly, the feasible set is finite and it is possible to compute the optimal bit loading numerically by trying out all possible candidates. This observation does however provide little insight into the overall behavior of the system. It is also questionable whether it is worth the computational burden to globally search all possible bit loading candidates. In order to gain more insight and to find heuristics for computing the bit rates more efficiently, the following section considers optimizing $\boldsymbol{b}$ while

the vectors $\boldsymbol{\xi}$ and $\boldsymbol{\sigma}$ remain fixed (recall that Chapter 8 was optimizing $\boldsymbol{\xi}$ and $\boldsymbol{\sigma}$ for a fixed $\boldsymbol{b}$). Then, later, Section 9.3 will be devoted to the joint optimization of all three vectors $\boldsymbol{\xi}$, $\boldsymbol{\sigma}$, and $\boldsymbol{b}$.

9.1 Continuous bit loading relaxation

Optimization of the discrete-valued bit loading is difficult in closed form. One way to approach this optimization problem is to relax the set of bit rates to the continuous domain (by ignoring the constraint (9.2a)), so that $\boldsymbol{b}$ can be analytically optimized. This leads to the continuous relaxation of Problem (9.1), where $\mathcal{B} = \mathbb{R}_+$. In order to specify constraint (9.1e), we assume for simplicity that the constellations are QAM. Then, using (7.9), the log-weights $\boldsymbol{w}$ depend on the bit allocations as

$$e^{\boldsymbol{w}} = \boldsymbol{d}(\boldsymbol{D}_w) = e^{\boldsymbol{b}} - \mathbf{1}, \tag{9.3}$$

where the unit of the rate has been changed to nats (rather than bits) in order to simplify the notation below. For a given $\boldsymbol{\xi}$ and $\boldsymbol{\sigma}$, using weights defined by (9.3), the continuous relaxation of Problem (9.1) with $\boldsymbol{\xi}$, $\boldsymbol{\sigma}$ fixed, is then formulated as

$$\underset{b_1,\ldots,b_N}{\text{minimize}} \quad \sum_{i=1}^{N} (e^{b_i} - 1)^p e^{-p\xi_i}, \tag{9.4a}$$

$$\text{subject to} \quad \sum_{i=1}^{N} b_i = R, \tag{9.4b}$$

$$b_i \geq 0 \quad \forall\, i, \tag{9.4c}$$

where we use the fact that $\|\cdot\|_p$ and $\|\cdot\|_p^p$ are minimized simultaneously[1]. Note (by inspection) that the problem is convex.

Theorem 9.1.1 *The optimum bit allocation in* (9.4) *is given by*

$$b_i = g(\nu + \xi_i) \quad \forall\, i = 1, ..., N, \tag{9.5}$$

where the function $g(x)$ *is defined as*

$$\begin{cases} x = (1 - p^{-1}) \log\left(1 - e^{-g(x)}\right) + g(x), & \text{if} \quad p > 1 \\ g(x) = (x)^+, & \text{if} \quad p = 1 \end{cases}, \tag{9.6}$$

and where ν *is chosen so that*

$$\sum_{i=1}^{N} g(\nu + \xi_i) = R. \tag{9.7}$$

[1]The infinity norm is not well defined at this point and has to be treated separately. The conclusions of the following discussion are however valid for the infinity norm as well.

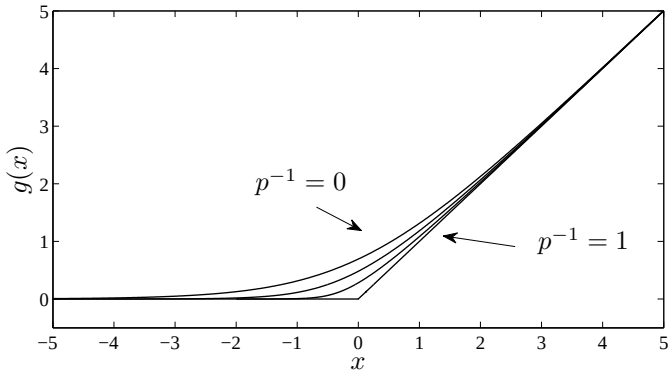

Figure 9.1: The $g(x)$ function for different levels of p.

Proof: In the following proof we assume $p \in (1, \infty)$. The proofs for $p = 1$ and $p \to \infty$ are similar but needs to be treated separately and are given in Appendix 9.3.2. Disregarding, for a while, the constraint that $\boldsymbol{b}$ must be positive; minimizing the Lagrangian cost function of (9.4) yields the optimal solution to the problem as

$$p\, e^{b_i}(e^{b_i} - 1)^{p-1} e^{-p\xi_i} = \theta \quad \forall\, i = 1, ..., N, \tag{9.8}$$

where θ is the non-negative dual variable such that constraint (9.4b) is satisfied. Equation (9.8) contains multiple roots. However, if there exists a root with strictly positive b_i's, then it must also be a global optimum to the convex problem (9.4) (a convex problem does not have local optima unless they also are global optima).

Taking the logarithm of (9.8) and performing some rearrangements yields

$$f(b_i) \triangleq (1 - p^{-1}) \log(1 - e^{-b_i}) + b_i = p^{-1} \log(\theta\, p^{-1}) + \xi_i. \tag{9.9}$$

Note that the function $f(b_i)$ is real valued when all b_i's are positive. By inspection, $f(b_i)$ is strictly increasing, concave, and it maps the set $(0, \infty)$ to the set $(-\infty, \infty)$. Because $f(b_i)$ is strictly increasing and concave, the inverse function $g(x)$ exists and is strictly increasing and convex. Figure 9.1 shows the function $g(x)$ for different p's. Following (9.9), the function $g(x)$ must satisfy

$$x = (1 - p^{-1}) \log\left(1 - e^{-g(x)}\right) + g(x). \tag{9.10}$$

This implies that for any vector $\boldsymbol{\xi}$ there exists one and only one solution to (9.8) with strictly positive bit rates, $\boldsymbol{b}$, given by

$$b_i = g(\nu + \xi_i) \quad \forall\, i = 1, ..., N, \tag{9.11}$$

where $\nu \triangleq p^{-1} \log(\theta\, p^{-1})$. □

Given a rate vector $\boldsymbol{\xi}$, equations (9.5) and (9.7) uniquely determines the optimal bit loading (as well as ν). These equations will later be applied to eliminate $\boldsymbol{b}$ from the joint optimization problem.

The observant reader may have noticed that the optimal relaxed bit loading will never be exactly zero on any subchannel. Instead of switching off weak subchannels with zero bit loading, it turns out that it is more favorable to use an infinitesimal (positive) bit rate. Of course there is no such thing as infinitesimal bit rates in practice, it is worth to recall that the relaxation is merely a tool that we can use to obtain practically implementable bit-loading candidates by means of rounding. Fortunately, as will be shown in Section 9.4, the impact that a low-rate subchannel has on the rest of the system is limited, i.e., the performance will remain close to optimal if we turn off the low-rate subchannels. After the bit loading, these low-rate subchannels will in any case be rounded to zero when we apply finite constellations. The next section contains a comment on the sensitivity of the relaxed optimum towards rounding.

9.2 Rounded bit loading

In practice, arbitrary real-valued bit rates are not implementable and the impact of rounding or quantization of the bit rates has to be considered. Assume $\tilde{\boldsymbol{b}}$, is the optimal solution to the relaxed bit loading problem for some given rate vector $\boldsymbol{\xi}$, and assume that $\boldsymbol{b}'$ is the rounded or quantized version of $\tilde{\boldsymbol{b}}$ such that the sum rate is R. Denote the logarithm of the objective function (9.1a) as

$$J(\boldsymbol{b},\boldsymbol{\xi}) = p^{-1}\log\Big(\sum_{i=1}^{N}(e^{b_i}-1)^p e^{-p\xi_i}\Big), \tag{9.12}$$

where we have chosen the weights according to (9.3). The first order Taylor expansion of (9.12) around the optimal bit loading $\tilde{\boldsymbol{b}}$ yields

$$J(\boldsymbol{b}',\boldsymbol{\xi}) \approx J(\tilde{\boldsymbol{b}},\boldsymbol{\xi}) + \frac{\sum_{i=1}^{N}(e^{\tilde{b}_i}-1)^{p-1}e^{\tilde{b}_i}e^{-p\,\xi_i}\,\delta_i}{e^{p\,J(\tilde{\boldsymbol{b}},\boldsymbol{\xi})}}, \tag{9.13}$$

where $\boldsymbol{\delta} = \boldsymbol{b}' - \tilde{\boldsymbol{b}}$. Now, using (9.8), and assuming both $\boldsymbol{b}'$ and $\tilde{\boldsymbol{b}}$ satisfy (9.4b) so that $\mathbf{1}^{\mathrm{T}}\boldsymbol{\delta} = 0$, the first order term in the expansion sums to zero

$$J(\boldsymbol{b}',\boldsymbol{\xi}) \approx J(\tilde{\boldsymbol{b}},\boldsymbol{\xi}) + \frac{\theta\,\sum_{i=1}^{N}\delta_i}{p\,e^{p\,J(\tilde{\boldsymbol{b}},\boldsymbol{\xi})}} = J(\tilde{\boldsymbol{b}},\boldsymbol{\xi}). \tag{9.14}$$

This result indicates that rounding of the optimal relaxed bit loading can be performed without too much loss in performance, although it is not clear how to quantify the loss. In the following section the loss is quantified for the joint optimum by making a distinction between low-rate and high-rate subchannels.

9.3 Joint optimization of bit loading and filters

Now that we know how to obtain the optimal DF filters (via $\boldsymbol{\xi}$ and $\boldsymbol{\sigma}$) for a given bit allocation (cf. Chapter 8), and how to optimize the bit allocation $\boldsymbol{b}$ given vectors $\boldsymbol{\xi}$ and $\boldsymbol{\sigma}$, our next step is to combine these results into a joint optimization problem.

9.3.1 The bit-loading optimized objective

The optimal relaxed bit allocation from Theorem 9.1.1 depends on the rate vector $\boldsymbol{\xi}$. By inserting the optimal bit loading into the refined transceiver problem (8.15), we obtain a new objective with a dependence on $\boldsymbol{\xi}$ that is not as easily characterized as before. This section analyzes the behavior with respect to $\boldsymbol{\xi}$ of such a bit-loading optimized objective. As it turns out, even though the dependency on $\boldsymbol{\xi}$ is complicated, it is still possible to determine the optimal $\boldsymbol{\xi}$ as a function of $\boldsymbol{\sigma}$.

The optimal relaxed bit-loading vector $\tilde{\boldsymbol{b}}$, obtained from (9.4) in the original objective, yields the following objective function

$$J(\xi_1, ..., \xi_N) = p^{-1} \log \Big(\sum_{i=1}^{N} (e^{\tilde{b}_i} - 1)^p e^{-p\xi_i} \Big), \tag{9.15}$$

where the logarithm is introduced for later mathematical simplicity. The optimal bit allocation must satisfy (9.5), which can be reformulated to[2]

$$e^{p\nu + p\xi_i} = \big(e^{g(\nu+\xi_i)} - 1\big)^p \big(1 - e^{-g(\nu+\xi_i)}\big)^{-1}, \tag{9.16}$$

and then, using $\tilde{b}_i = g(\nu + \xi_i)$ as

$$(e^{\tilde{b}_i} - 1)^p e^{-p\xi_i} = e^{p\nu}\big(1 - e^{-g(\nu+\xi_i)}\big) \quad \forall\, i, \tag{9.17}$$

we obtain the bit-loading optimized cost function without the vector $\tilde{\boldsymbol{b}}$ as

$$J(\xi_1, ..., \xi_N) = \ \nu + p^{-1} \log \sum_{i=1}^{N} \Big(1 - e^{-g(\nu+\xi_i)}\Big), \tag{9.18}$$

where ν is chosen such that

$$\sum_{i=1}^{N} g(\nu + \xi_i) = R. \tag{9.19}$$

Note that (9.18) is also valid for the ∞-norm when $p^{-1} = 0$.

Using the new bit-optimized cost function, the remaining optimization problem (that determines the DF filters) is

$$\begin{aligned} &\underset{\boldsymbol{\xi},\boldsymbol{\sigma}}{\text{minimize}} && J(\boldsymbol{\xi}) && \text{(9.20a)} \\ &\text{subject to} && \boldsymbol{\xi} \preceq \log(\boldsymbol{\Lambda}_{\boldsymbol{H}}^2 \boldsymbol{\sigma} + \mathbf{1}), && \text{(9.20b)} \\ & && \boldsymbol{\sigma} \geq 0, \quad \mathbf{1}^{\mathrm{T}} \boldsymbol{\sigma} \leq P. && \text{(9.20c)} \end{aligned}$$

[2]Compare with (9.8).

Although the problem is non-convex and perhaps difficult to solve, it turns out the cost function is symmetric and concave which enables us to solve at least parts of the problem with relative ease.

Theorem 9.3.1 *The function $J(\boldsymbol{\xi})$ is Schur-concave with respect to $\boldsymbol{\xi}$.*

Proof: See Appendix 9.C. □

A direct consequence of Theorem 9.3.1 is the following important corollary

Corollary 9.3.2 *Orthogonal SVD-based transmission with no decision feedback is always an optimal solution to the decision feedback problem, given that the optimal relaxed bit loading is used.*

Proof: Because the objective is Schur-concave (see Appendix 9.A for definition), the optimal vector $\boldsymbol{\xi}$ must satisfy the majorization constraint with equality (cf. [JB06]), i.e. we have $\boldsymbol{\xi} = \log(\mathbf{1} + \boldsymbol{\Lambda}_{\boldsymbol{H}}^2 \boldsymbol{\sigma})$. This means that $\boldsymbol{V}_{\boldsymbol{F}}$ in (8.9) can be chosen as the identity matrix so that the subchannels are orthogonalized. Orthogonal subchannels implies that $\boldsymbol{L}$ is diagonal and that the optimal feedback matrix $\boldsymbol{B}$ is zero (see Section 8.2). □

The remaining problem of computing the optimal power load, $\boldsymbol{\sigma}$, is in general a problem with a non-trivial solution. However, the result above shows that it suffices to use SVD-based bit and power loading schemes to compute a close-to-optimal bit loading, e.g., by using the gap approximation (see Chapter 2.6.1). For the infinity norm it is actually possible to obtain a solution for the optimized power.

9.3.2 Joint bitloading–power optimization: ∞-norm

In this section we derive the solution for the infinity norm. The problem of computing the optimal power load $\boldsymbol{\sigma}$ is

$$\underset{\boldsymbol{\sigma}}{\text{minimize}} \qquad J\big(\log(\boldsymbol{\Lambda}_{\boldsymbol{H}}^2 \boldsymbol{\sigma} + \mathbf{1})\big) \tag{9.21a}$$

$$\text{subject to} \qquad \boldsymbol{\sigma} \geq 0, \quad \mathbf{1}^{\mathrm{T}} \boldsymbol{\sigma} \leq P, \tag{9.21b}$$

where ν is given by the equation

$$\sum_{i=1}^{N} \log\big((\lambda_i \sigma_i + 1)e^{\nu} + 1\big) = R. \tag{9.22}$$

First we need to establish that the global optimum is solvable my means of differentiation, i.e. we need to know that the cost function is convex.

Theorem 9.3.3 *For the case $p^{-1} = 0$, the Hessian of the function*

$$J\big(\log(\boldsymbol{\Lambda}_{\boldsymbol{H}}^2 \boldsymbol{\sigma} + \mathbf{1})\big), \tag{9.23}$$

with respect to $\boldsymbol{\sigma}$ is positive semi-definite.

Proof: See Appendix 9.D. □

Hence, the global optimum can be computed by solving the Karush-Kuhn-Tucker (KKT) conditions of the problem. It turns out the solution is easier to obtain if we reformulate the problem as follows. For the infinity norm $p^{-1} = 0$, from (9.18) we have $J\big(\log(\mathbf{\Lambda}_{\boldsymbol{H}}^2\boldsymbol{\sigma} + \mathbf{1})\big) = \nu$. This means that problem (9.21) is equivalent to

$$\underset{\boldsymbol{\sigma}}{\text{maximize}} \qquad \sum_{i=1}^{N} \log\big((\lambda_i\sigma_i + 1)e^{\nu} + 1\big) \tag{9.24a}$$

$$\text{subject to} \qquad \boldsymbol{\sigma} \geq 0, \quad \mathbf{1}^{\mathrm{T}}\boldsymbol{\sigma} \leq P, \tag{9.24b}$$

Solving the KKT conditions yields the optimal $\boldsymbol{\sigma}$ and ν as the solution to the following equations

$$\sigma_i = \big(\mu - (1 + e^{-\nu})\,\lambda_i^{-1}\big)^{+}, \quad \forall\, i = 1, ..., N, \tag{9.25a}$$

$$R = \sum_{i=1}^{N} \log\big(e^{\nu}(\lambda_i\sigma_i + 1) + 1\big), \tag{9.25b}$$

$$P = \mathbf{1}^{\mathrm{T}}\boldsymbol{\sigma}, \tag{9.25c}$$

where μ is a dual variable such that the power constraint is satisfied. The optimal bit loading is, after some derivations,

$$b_i = \max\big(\log(\lambda_i) + \alpha, \log(e^{\nu} + 1)\big), \tag{9.26}$$

where α ensures that the total bit rate is R nats. Again we note the fact that zero power on a substream gives a non-zero bit load

$$b_{min} = \log(e^{\nu} + 1). \tag{9.27}$$

Two reasons can explain this somewhat counterintuitive result. The first is that the objective, to minimize the maximum SER, is not applicable when operating in the high error-rate regime. To see this, consider the case when the data rate assigned to a particular subchannel is very high, with the corresponding SER above 0.5. In such case the objective can always be reduced by throwing away one bit (or equivalently, put it on a subchannel with zero power) since the single bit has a SER of 0.5. Table 9.1 illustrates this example in more detail. Bit Loading 1 puts all eight bits on the only connected sub channel, while Bit Loading 2 throws away two bits. It is clear that the min-max SER strategy is to throw away two bits, although this is not the minimum BER strategy (which perhaps is more relevant for the high error rate regime). The other reason for a non-zero b_{min} is that the minimum distance formula (7.8) is exact only for QAM constellations with an even integer bit rate and not well defined for bit rates close to zero.

	Bit Loading 1	Bit Loading 2
$\boldsymbol{b}$	(8; 0; 0)	(6; 1; 1)
SER	(0.758; 0; 0)	(0.348; 0.5; 0.5)
BER	(0.142; 0; 0)	(0.064; 0.5; 0.5)
<u>BER</u>	0.142	0.173

Table 9.1: Bit loading on three additive white Gaussian noise channels with Gray coded constellations. The first subchannel has 15 dB SNR, while the other two are unconnected (SNR: $-\infty$ dB).

In practice, however, for practically useful symbol error rates, the value of $b_{min} \ll 1$ bits and will be rounded to zero rather than to QPSK. When rounding b_{min} to zero, the optimal relaxed bit loading is

$$b_i = \big(\log(\lambda_i) + \alpha\big)^+, \tag{9.28}$$

which is identical to what is attained using the well known gap approximation[3] for bit loading (in combination with optimal power loading) of orthogonal substreams (cf. [PB05] or Chapter 2). This confirms the conclusion drawn from Corollary 9.3.2 that one should use classical bit and power loading schemes designed for orthogonal subchannels even if a DF filter is available. If it should happen that b_{min} is not rounded to zero, then there are reasons to believe the system is operating in a region for which the min-max SER is no longer a good objective.

9.4 Turning off low-rate subchannels

Theorem 9.3.1 relies on the continuous relaxation, and the behavior of $J(\boldsymbol{\xi})$ for discrete constellation sets is not clear at this point. On the other hand, as was shown in Section 9.2, small perturbations of the optimal relaxed bit load will not significantly alter the value of the cost function. So, rounding the optimal bit loading should still remain close to optimal. In the this section, an upper bound on the loss due to rounding of the bit rates is derived.

Essentially, rounding the bit loading corresponds to turning off low-rate subchannels and slightly perturbing the bit rates on the remaining high-rate subchannels. We will show that the loss by turning off low-rate subchannels is relatively small and then, that a system with no active low-rate subchannels is insensitive to reallocations of the bit loading.

An interesting property of $g(x)$ is that its asymptotes[4] coincide with the function $(x)^+$. Therefore, by analyzing (9.18), (9.19), we see that weak subchannels with

[3] In fact, by using the gap approximation $b_i = \log(e^{\nu_{\mathrm{GAP}}}(e^{\xi_i} - 1) + 1)$ instead of the min-max SER bit loading (9.5), it can be shown (similar to Theorem 9.3.1) that the objective $\nu_{\mathrm{GAP}}(\boldsymbol{\xi})$ is also Schur-concave. Thus Corollary 9.3.2 holds for this bit loading strategy as well.

[4] First note that the range of $g(x)$ is non-negative, then from (9.6) we see that $g(x) \gg 1 \Rightarrow x \approx g(x)$ and $g(x) \approx 0 \Rightarrow x \approx (1 - p^{-1}) \log g(x)$.

values of x that are negative or close to zero will have almost no impact on ν or on $J(\xi_1, ..., \xi_N)$. These subchannels can consequently be turned off at a very low cost in terms of performance. To formalize this, assume that $\boldsymbol{\xi}$ is decreasing and that all $N - \tilde{N}$ weakest subchannels with indices $i > \tilde{N}$ are turned off. This will result in a new dual variable $\tilde{\nu}$ and cost function $\tilde{J}(\xi_1, ..., \xi_{\tilde{N}})$ as

$$\tilde{J}(\xi_1, ..., \xi_{\tilde{N}}) = \tilde{\nu} + p^{-1} \log \sum_{i=1}^{\tilde{N}} \left(1 - e^{-g(\tilde{\nu}+\xi_i)}\right), \tag{9.29}$$

$$\sum_{i=1}^{\tilde{N}} g(\tilde{\nu} + \xi_i) = R. \tag{9.30}$$

The following theorem quantifies the loss.

Theorem 9.4.1 *The loss when turning off the $N - \tilde{N}$ weakest subchannels can be upper bounded as*

$$\tilde{J}(\xi_1, ..., \xi_{\tilde{N}}) - J(\xi_1, ..., \xi_N) \leq \frac{\sum_{i=\tilde{N}+1}^{N} g(\nu + \xi_i)}{\tilde{N} - \sum_{i=1}^{\tilde{N}} \frac{1-p^{-1}}{e^{g(\nu+\xi_i)}-p^{-1}}}. \tag{9.31}$$

Proof: See Appendix 9.E. □

In order to get a sense of how this bound behaves, denote the sum rate of the truncated subchannels

$$\Delta_R = \sum_{i=\tilde{N}+1}^{N} g(\nu + \xi_i). \tag{9.32}$$

Assuming the active subchannels satisfy $e^{g(\nu+\xi_i)} \gg 1$, then the denominator in (9.31) can be approximated with $\tilde{N}$, and the bound becomes

$$\tilde{J} - J \leq \frac{\Delta_R}{R} \cdot \frac{R}{\tilde{N}}. \tag{9.33}$$

As an example, typical figures for Δ_R/R could be on the order of 10% while the average data rate per active subchannel is typically less than, lets say, 3 nats. This would correspond to a maximum loss of around 1 dB.

The next step is to see how the cost function behaves when all low-rate subchannels have been turned off. Given that all subchannels are high rate we can apply $e^{g(\tilde{\nu}+\xi_i)} \gg 1$ to the definition (9.6) and obtain

$$g(\tilde{\nu} + \xi_i) \approx \tilde{\nu} + \xi_i. \tag{9.34}$$

By applying the asymptote to (9.30), the cost function (9.29) tends to

$$\tilde{J}(\xi_1, ..., \xi_{\tilde{N}}) \approx \frac{R}{\tilde{N}} - \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \xi_i + p^{-1} \log(\tilde{N}). \tag{9.35}$$

Given that the majorization constraint (9.20b) is satisfied, the following equality holds

$$\sum_{i=1}^{\tilde{N}} \xi_i = \sum_{i=1}^{\tilde{N}} \log(\lambda_i \sigma_i + 1), \tag{9.36}$$

and we can eliminate $\boldsymbol{\xi}$ completely from (9.35). Interestingly, any $\boldsymbol{\xi}$ that satisfies (9.20b) can be used and still be optimal. Since there is a direct relation between the optimal $\boldsymbol{b}$ and $\boldsymbol{\xi}$, this result implies that we can redistribute the bit allocations at a very low cost, provided the resulting $\boldsymbol{\xi}$ satisfies (9.20b) and the data rates on the active subchannels remain sufficiently high.

The results in this section predict very limited losses when rounding the relaxed bit loading. This fact is further motivated by the numerical results in Section 9.6 where almost identical performance of the truly optimal bit loading (achieved by a global search) and the rounded optimal relaxed bit loading is shown.

9.5 Transmission schemes

Due to the potential high complexity of the truly optimal joint bit loading and filter design, this section defines (in addition to the optimal design) three suboptimal schemes: Two of which, in theory, should perform very close to optimal.

9.5.1 Optimal design

This transmission scheme is optimal in terms of (7.15). The strategy is to exhaustively search all combinations of bit loading allocations. For each bit loading candidate, optimize the rate vector, $\boldsymbol{\xi}$, and the power loading vector, $\boldsymbol{\sigma}$, by solving Problem (8.16). Compute the weighted MSEs and use the bit loading with the least weighted MSE.

9.5.2 Suboptimal bit loading

As Theorem 9.3.1 shows, the optimal relaxed bit loading allow us to make the DF optimized system orthogonal. In [PB05], the so called gap approximation was used for determining the constellations of an orthogonal system (see also Section 2.6.1). The gap approximation is close to optimal for the orthogonal system, and since the optimal bit loading with DF results in an orthogonal system it must be approximately optimal in this case as well.

Given that the power as well as the bit loading has been optimized (as was done in Section 2.6.1), the gap approximation leads to the following bit loading

$$b_i = 2 \left\lfloor \frac{1}{2} \log_2(\lambda_i) + \alpha \right\rfloor^+, \tag{9.37}$$

where α is a constant such that $\sum_i b_i = R$. The precoder, forward filter, and feedback filter are then optimized for this particular bit loading.

9.5.3 Orthogonal design

As was shown in Section 9.3, the optimal relaxed bit loading allows us to use orthogonal subchannels. It was also shown that the first order Taylor expansion around the optimal relaxed bit loading is constant. Hence, an optimal design under the constraint that the subchannels are forced to be orthogonal should perform almost as good as the two schemes above.

Use the gap approximation to compute the bit rates, then design the optimal *orthogonalizing* precoder and forward filter for this particular bit loading. That is, design the optimal precoder such that the interference among the subchannels is zero. Since the subchannels are orthogonal for this scheme, the optimal feedback matrix will be zero.

9.5.4 Equal rate design

The bit rates are distributed uniformly among all available subchannels. Again, the precoder, the forward filter, and the feedback filter are subsequently optimized for this particular bit loading.

9.6 Numerical results

In this section we numerically compare the schemes that was introduced in Section 9.5. For simplicity, only the infinity norm has been considered as cost function. Figure 9.2 shows a comparison of the scheme over an 8×8 Rayleigh-fading MIMO channel. The data rate is set to 24 bits per channel use. The optimal design and the suboptimal bit loading design have almost identical performance. This confirms that the rounding of the bits does not affect the overall performance significantly, and the DF filters compensate for small deviations from the optimal bit loading. The orthogonal design performs only slightly worse, which is an indication that if the appropriate bit loading is used, the importance of DF is very limited. The final scheme, equal bit loading, shows that the bit loading is important for achieving optimal performance. The difference in performance is however less than one dB and this indicates that DF can partly compensate for suboptimal bit loading. Figure 9.3 and Figure 9.4 illustrate similar results: In Figure 9.3, a 6×6 Rayleigh-fading channel was simulated with a data rate at 18 bits per channel use. Figure 9.4 similarly illustrates transmission over a 4×4 Rayleigh-fading channel was simulated with a data rate at 12 bits per channel.

9.7 Conclusions

In this part of the thesis, we considered the problem of joint optimization of the bit loading, precoder, and receiver filters for a DF-detection system. It was shown that minimizing the probability of detection error can be translated into minimizing a weighted p-norm of the MSEs. Then, by fixing the bit loading, it was shown

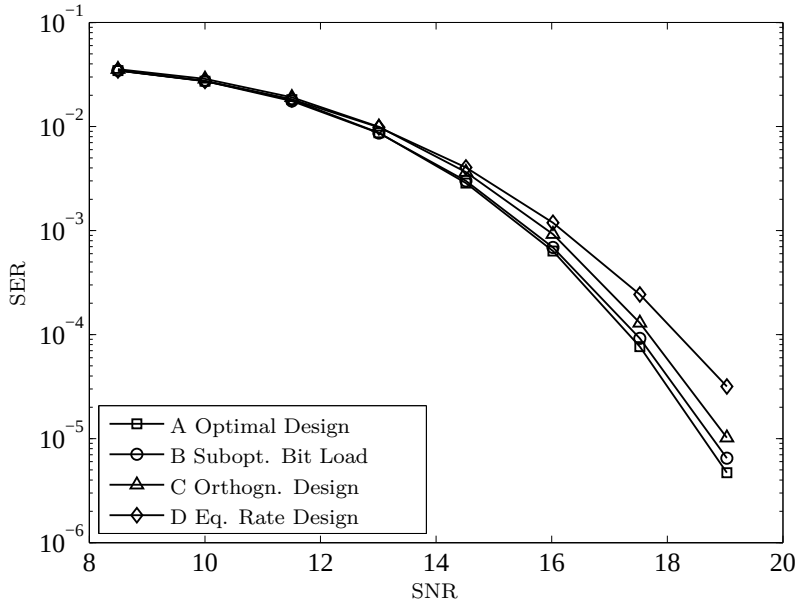

Figure 9.2: Monte Carlo simulations of an 8×8 MIMO system. The data rate is set to 24 bits, and the cost function is the infinity norm.

that the problem of optimizing the precoder and receiver filters may be reduced to a convex optimization problem that is easy to solve numerically. Due to the low computational complexity of the problem, the task of jointly optimizing the bit loading and filters by exhaustively searching through all possible bit-loading candidates becomes a feasible option in practice.

In another approach to the same problem, by instead fixing the DF filters, we derived the optimal bit loading by relaxing the integer constraints on the subchannel bit rates. It was shown that this optimum is insensitive towards small deviations in the bit loading. When combining the relaxed bit loading with filter optimization, we showed that it is optimal to use orthogonal non-interfering subchannels. Therefore, by jointly optimizing bit loading and filters, the DF part of the receiver becomes superfluous. That said, another conclusion is that the DF receiver makes the system more robust towards rounding of the bit loading. These results were illustrated numerically by comparisons between the truly optimal solution and various suboptimal transmission strategies.

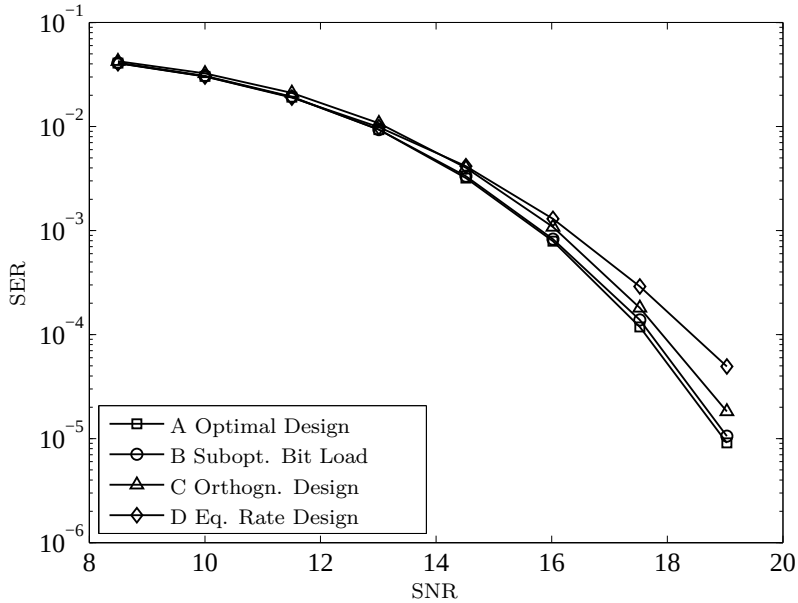

Figure 9.3: Monte Carlo simulations of a 6×6 MIMO system. The data rate is set to 18 bits, and the cost function is the infinity norm.

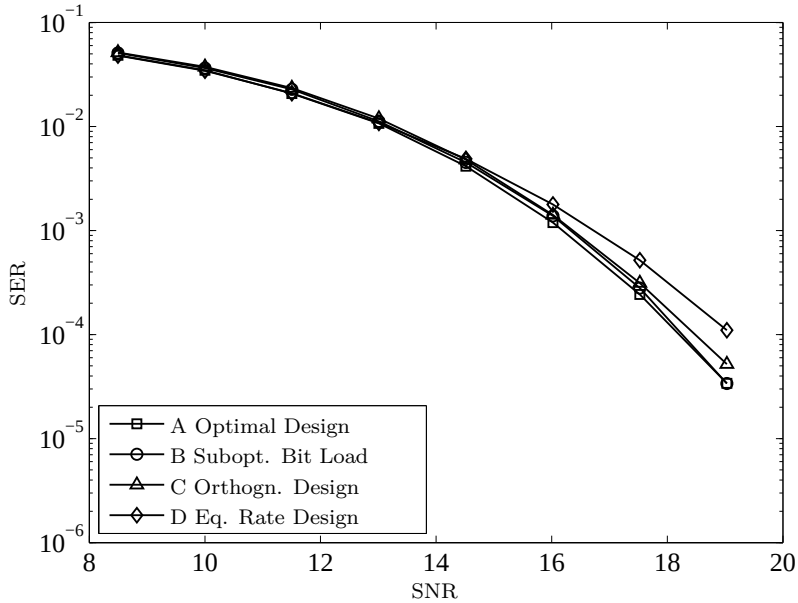

Figure 9.4: Monte Carlo simulations of a 4×4 MIMO system. The data rate is set to 12 bits, and the cost function is the infinity norm.

Appendix 9.A Definitions from majorization theory

Denote $x_{[1]}, x_{[2]}, ..., x_{[N]}$ as the monotonic rearrangement of a vector $\boldsymbol{x}$ such that $x_{[1]} \geq x_{[2]} \geq ... \geq x_{[N]}$. For two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, additive majorization is defined as

$$\boldsymbol{x} \preceq \boldsymbol{y} \;\Leftrightarrow\; \begin{cases} \sum_{i=1}^{k} x_{[i]} \leq \sum_{i=1}^{k} y_{[i]} \;\forall\, k = 1, ..., N-1 \\ \sum_{i=1}^{N} x_{[i]} = \sum_{i=1}^{N} y_{[i]} \end{cases}.$$

Similarly, multiplicative majorization is defined as

$$\boldsymbol{x} \preceq_{\times} \boldsymbol{y} \;\Leftrightarrow\; \begin{cases} \prod_{i=1}^{k} x_{[i]} \leq \prod_{i=1}^{k} y_{[i]} \;\forall\, k = 1, ..., N-1 \\ \prod_{i=1}^{N} x_{[i]} = \prod_{i=1}^{N} y_{[i]} \end{cases}.$$

A function $f(\boldsymbol{x})$ is said to be Schur-convex if

$$\boldsymbol{x} \preceq \boldsymbol{y} \quad \Longrightarrow \quad f(\boldsymbol{x}) \leq f(\boldsymbol{y}), \tag{9.38}$$

similarly it is defined Schur-concave if

$$\boldsymbol{x} \preceq \boldsymbol{y} \quad \Longrightarrow \quad f(\boldsymbol{x}) \geq f(\boldsymbol{y}). \tag{9.39}$$

Multiplicative Schur-convex/concave functions are defined in a similar fashion using $\preceq_{\times}$ instead of $\preceq$. For a more complete introduction to majorization theory, please see [JB06].

Appendix 9.B Extended proof of Theorem 9.1.1

In this appendix we prove Theorem 9.1.1 for the two special cases $p = 1$ and $p \to \infty$. The first case, $p = 1$, corresponds to the following optimization problem

$$\underset{b_1,...,b_N}{\text{minimize}} \qquad \sum_{i=1}^{N} (e^{b_i} - 1) e^{-\xi_i}, \tag{9.40a}$$

$$\text{subject to} \qquad \sum_{i=1}^{N} b_i = R, \tag{9.40b}$$

$$b_i \geq 0 \quad \forall\, i. \tag{9.40c}$$

Solving the KKT conditions results in

$$\begin{cases} e^{b_i - \xi_i} = \theta & \text{if } b_i > 0 \\ b_i = 0 & \text{otherwise} \end{cases}, \tag{9.41}$$

where θ is a positive dual such that

$$\sum_{i=1}^{N} b_i = R. \tag{9.42}$$

Note that (9.41) is equivalent to

$$b_i = (\nu + \xi_i)^+, \qquad \nu = \log(\theta), \tag{9.43}$$

which is exactly the solution given by Theorem 9.1.1.

In the second case, $p \to \infty$, we must use the cost function on the original form $\|\cdot\|_p$, rather than $\|\cdot\|_p^p$. Hence, the problem is

$$\underset{b_1,\ldots,b_N}{\text{minimize}} \qquad \max_i (e^{b_i} - 1)e^{-\xi_i}, \tag{9.44a}$$

$$\text{subject to} \qquad \sum_{i=1}^{N} b_i = R, \tag{9.44b}$$

$$b_i \geq 0 \quad \forall\, i. \tag{9.44c}$$

By inspection, the solution is given by

$$e^{b_i} - 1 = e^{\xi_i}\theta, \tag{9.45}$$

where, again, θ is a positive dual such that the sum rate is R. Reformulating the expression yields

$$b_i = \log(1 + e^{-\xi_i - \nu}) + \xi_i + \nu, \tag{9.46}$$

where ν is chosen such that the sum rate is R. This solution is also equivalent to the solution given by Theorem 9.1.1. □

Appendix 9.C Proof of Theorem 9.3.1

First we will derive the second order derivative of the cost function (9.18), then, using the second order derivatives, we will show that the cost function is concave.

Let $\phi = p^{-1} \in [0, 1]$. The derivative of $g(x)$ is

$$g' = \frac{\partial g}{\partial x} = \frac{e^g - 1}{e^g - \phi}. \tag{9.47}$$

Rearranging the equation yields

$$\phi g' e^{-g} = g' + e^{-g} - 1. \tag{9.48}$$

Define $g_i = g(\nu + \xi_i)$, and $g_i' = g'(\nu + \xi_i)$, then the derivative with respect to ξ_n is

$$\frac{\partial g_i}{\partial \xi_n} = \frac{\partial g(\nu + \xi_i)}{\partial \xi_n} = \left(\frac{\partial \nu}{\partial \xi_n} + \delta_{n,i}\right) g_i', \tag{9.49}$$

where $\delta_{n,i} = 1$ if $n = i$, and zero otherwise. Differentiating (9.19) with respect to ξ_n results in

$$\frac{\partial \nu}{\partial \xi_n} = -\frac{g_n'}{\sum_i g_i'}. \tag{9.50}$$

Using (9.49), then (9.48), and finally (9.50), the derivative of (9.18) with respect to ξ_n is

$$\begin{aligned}\frac{\partial J}{\partial \xi_n} &= \frac{\partial \nu}{\partial \xi_n} + \frac{\sum_i \phi e^{-g_i} g_i'}{\sum_i 1 - e^{-g_i}} \frac{\partial \nu}{\partial \xi_n} + \frac{\phi e^{-g_n} g_n'}{\sum_i 1 - e^{-g_i}} \\ &= \frac{\sum_i g_i'}{\sum_i 1 - e^{-g_i}} \frac{\partial \nu}{\partial \xi_n} + \frac{g_n' + e^{-g_n} - 1}{\sum_i 1 - e^{-g_i}} \\ &= -\frac{1 - e^{-g_n}}{\sum_i 1 - e^{-g_i}}\end{aligned} \tag{9.51}$$

The second order derivative of $J(\cdot)$ with respect to ξ_n, ξ_m, is given by

$$\begin{aligned}\frac{\partial^2 J}{\partial \xi_n \partial \xi_m} = &-\frac{e^{-g_n}}{\sum_i 1 - e^{-g_i}} \frac{\partial g_n}{\partial \xi_m} + \\ &+ \frac{1 - e^{-g_n}}{(\sum_i 1 - e^{-g_i})^2} \Big(\sum_i e^{-g_i} \frac{\partial g_i}{\partial \xi_m} \Big).\end{aligned} \tag{9.52}$$

Using (9.49) and (9.50) we can obtain

$$\frac{\partial g_n}{\partial \xi_m} = \sqrt{g_n'} \left(\delta_{n,m} - \frac{\sqrt{g_n'}\sqrt{g_m'}}{\sum_j g_j'} \right) \sqrt{g_m'}. \tag{9.53}$$

In order to show that $J(\boldsymbol{\xi})$ is concave, we will compute the Hessian matrix. To do that, we first introduce the following diagonal matrices

$$[\boldsymbol{A}]_{i,i} = g_i', \qquad [\boldsymbol{B}]_{i,i} = 1 - e^{-g_i}. \tag{9.54}$$

Note that since $g \geq 0$, and $g' \geq 0$, both $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive semi-definite (PSD). With these definitions we can define a matrix $\boldsymbol{G}$ as

$$\begin{aligned}&[\boldsymbol{G}]_{n,m} = \frac{\partial g_n}{\partial \xi_m} \quad \Longrightarrow \\ &\boldsymbol{G} = \boldsymbol{A}^{1/2} \left(\boldsymbol{I} - \frac{\boldsymbol{A}^{1/2} \mathbf{1} \mathbf{1}^{\mathrm{T}} \boldsymbol{A}^{1/2}}{\mathbf{1}^{\mathrm{T}} \boldsymbol{A} \mathbf{1}} \right) \boldsymbol{A}^{1/2},\end{aligned} \tag{9.55}$$

and then, using (9.52), we can derive the Hessian matrix of $J(\cdot)$ as

$$\boldsymbol{H} = -\frac{(\boldsymbol{I} - \boldsymbol{B})\boldsymbol{G}}{\mathbf{1}^{\mathrm{T}} \boldsymbol{B} \mathbf{1}} + \frac{\boldsymbol{B} \mathbf{1} \mathbf{1}^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{B}) \boldsymbol{G}}{(\mathbf{1}^{\mathrm{T}} \boldsymbol{B} \mathbf{1})^2}. \tag{9.56}$$

Because $\boldsymbol{G}^{\mathrm{T}} \mathbf{1} = \mathbf{0}$, the Hessian can be simplified as

$$\boldsymbol{H} = -\frac{\boldsymbol{C}\boldsymbol{G}}{\mathbf{1}^{\mathrm{T}} \boldsymbol{B} \mathbf{1}}. \tag{9.57}$$

where

$$\boldsymbol{C} \triangleq \boldsymbol{I} - \boldsymbol{B} + \frac{\boldsymbol{B}\mathbf{1}\mathbf{1}^{\mathrm{T}}\boldsymbol{B}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{B}\mathbf{1}}. \tag{9.58}$$

We will show in a few steps that this Hessian, $\boldsymbol{H}$, is a negative semi-definite matrix. As a reference regarding the various properties of PSD matrices we refer to [HJ85]. The center factor of $\boldsymbol{G}$ is a projection matrix (thus PSD), and consequently the entire matrix $\boldsymbol{G}$ is PSD. By inspection, the matrices $\boldsymbol{I} - \boldsymbol{B}$ and $\boldsymbol{B}\mathbf{1}\mathbf{1}^{\mathrm{T}}\boldsymbol{B}$ are both PSD, and because the sum of two PSD matrices is also PSD, $\boldsymbol{C}$ is PSD. The eigenvalues of the product of two PSD matrices are always real and non-negative, and consequently we know that the Hessian has non-positive real eigenvalues. Any real, symmetric matrix with non-positive real eigenvalues is negative semi-definite, hence the Hessian is negative semi-definite[5]. By inspection, the function $J(\cdot)$ is component-wise symmetric and thus, because it is jointly concave, it is also Schur-concave [JB06]. □

Appendix 9.D Proof of Theorem 9.3.3

Because $\boldsymbol{\Lambda}_{\boldsymbol{H}}^2\boldsymbol{\sigma} + \mathbf{1}$ is affine, it suffices to show that $J(\log \boldsymbol{\alpha})$ is convex with respect to $\boldsymbol{\alpha}$. The second order derivative of $J(\log \boldsymbol{\alpha})$ is

$$\frac{\partial^2 J(\log \boldsymbol{\alpha})}{\partial \alpha_i \partial \alpha_j} = \frac{1}{\alpha_i}\left(\frac{\partial^2 J(\boldsymbol{\xi})}{\partial \xi_i \partial \xi_j} - \delta_{i,j}\frac{\partial J(\boldsymbol{\xi})}{\partial \xi_i}\right)\frac{1}{\alpha_j}. \tag{9.59}$$

For the second order derivative to be positive definite we need to show that the matrix $\boldsymbol{E}$, defined as

$$[\boldsymbol{E}]_{i,j} = \frac{\partial^2 J(\boldsymbol{\xi})}{\partial \xi_i \partial \xi_j} - \delta_{i,j}\frac{\partial J(\boldsymbol{\xi})}{\partial \xi_i}, \tag{9.60}$$

is PSD. Introducing the matrix notation from Appendix 9.C yields

$$\boldsymbol{E} = \frac{1}{\mathbf{1}^{\mathrm{T}}\boldsymbol{B}\mathbf{1}}\big(\boldsymbol{B} - \boldsymbol{C}\boldsymbol{G}\big). \tag{9.61}$$

Using the assumption that $p^{-1} = 0$ we have $\boldsymbol{A} = \boldsymbol{B}$, $\boldsymbol{C} = \boldsymbol{I} - \boldsymbol{G}$, and thus

$$\begin{aligned}\boldsymbol{E} &= \frac{1}{\mathbf{1}^{\mathrm{T}}\boldsymbol{B}\mathbf{1}}\big(\boldsymbol{B} - \boldsymbol{G} + \boldsymbol{G}\boldsymbol{G}\big) \\ &= \frac{1}{\mathbf{1}^{\mathrm{T}}\boldsymbol{B}\mathbf{1}}\left(\frac{\boldsymbol{B}\mathbf{1}\mathbf{1}^{\mathrm{T}}\boldsymbol{B}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{B}\mathbf{1}} + \boldsymbol{G}\boldsymbol{G}\right),\end{aligned} \tag{9.62}$$

which is PSD by inspection. □

[5]Symmetry is perhaps not apparent from (9.57). However, all second order derivatives of $J(\cdot)$ are continuous and consequently we know that the Hessian matrix (9.57) is symmetric. The interested reader can alternatively show symmetry of (9.57) by applying (9.48), but this requires a few extra steps of derivations.

Appendix 9.E Proof of Theorem 9.4.1

Due to (9.19), the dual variable ν will inevitably be affected when reducing the number of subchannels. Denote the new dual variable $\tilde{\nu}$, and equation (9.19) gives

$$\sum_{i=1}^{\tilde{N}} g(\tilde{\nu}+\xi_i) = \sum_{i=1}^{N} g(\nu+\xi_i). \tag{9.63}$$

Since $g(x) \geq 0$ and $g'(x) \geq 0$, we have $\tilde{\nu} \geq \nu$. The convexity of $g(x)$ implies

$$g(\tilde{\nu}+\xi_i) - g(\nu+\xi_i) \geq g'(\nu+\xi_i)(\tilde{\nu}-\nu), \tag{9.64}$$

where the derivative is specified in (9.47). Apply (9.64) to (9.63) as

$$\tilde{\nu} - \nu \leq \frac{\sum_{i=\tilde{N}+1}^{N} g(\nu+\xi_i)}{\tilde{N} - \sum_{i=1}^{\tilde{N}} \frac{1-p^{-1}}{e^{g(\nu+\xi_i)}-p^{-1}}}. \tag{9.65}$$

Note that, because $\tilde{\nu} \geq \nu$, $g(x)$ is positive and increasing, and because $\xi_{\tilde{N}+1}, ..., \xi_N$ correspond to the weakest subchannels; the vectors

$$\begin{aligned} \tilde{\boldsymbol{b}} &= \big[\, g(\tilde{\nu}+\xi_1), ..., g(\tilde{\nu}+\xi_{\tilde{N}}), 0, ..., 0 \,\big]^{\mathrm{T}}, \\ \boldsymbol{b} &= \big[\, g(\nu+\xi_1), ..., g(\nu+\xi_N) \,\big]^{\mathrm{T}}, \end{aligned} \tag{9.66}$$

of length N satisfy $\boldsymbol{b} \preceq \tilde{\boldsymbol{b}}$. Since $\mathbf{1}^{\mathrm{T}} e^{-\boldsymbol{b}}$ is a Schur-convex function we therefore have $\mathbf{1}^{\mathrm{T}} e^{-\boldsymbol{b}} \leq \mathbf{1}^{\mathrm{T}} e^{-\tilde{\boldsymbol{b}}}$, and consequently

$$\log \sum_{i=1}^{\tilde{N}} \Big(1 - e^{-g(\tilde{\nu}+\xi_i)}\Big) \leq \log \sum_{i=1}^{N} \Big(1 - e^{-g(\nu+\xi_i)}\Big). \tag{9.67}$$

This together with (9.65) proves the theorem. □

Chapter 10

Skewed majorization

This chapter considers the optimization of Schur-convex objective functions given shifted or skewed majorization constraints that appeared in Chapter 8.

10.1 Introduction

A majorization inequality consists of a specific collection of inequalities and is used for comparing vectors. These inequalities appear, e.g., when comparing the diagonal elements of a positive semi-definite matrix with its singular values, or similarly, when comparing the Cholesky factor's diagonal elements with the singular values [HJ91].

Optimization problems constrained by a majorization inequality appears in various applications [PJ07, PCL03, SD08]. If the cost function can be characterized as either Schur-convex or Schur-concave (see Appendix 9.A), the optimal solution to the problem is given instantly according to rules of majorization theory [JB06]. In the case when the cost function is convex — but not necessarily Schur-convex — the solution can be calculated by numerical means using convex optimization [BV04].

In this chapter, we consider the minimization of a Schur-convex objective function, but where the majorization constraint is skewed or shifted linearly. More specifically; a skewed majorization constraint implies that the vector variable $\boldsymbol{z}$ must satisfy that $\boldsymbol{z}+\boldsymbol{w}$ is majorized by $\boldsymbol{c}$, where $\boldsymbol{w}$ and $\boldsymbol{c}$ are vector parameters to the problem. The vector $\boldsymbol{w}$ shifts or "skewes" the constraint, making conventional majorization theory not directly applicable. The skewed majorization problem can be posed in many different ways; in Appendix 10.A two alternative, albeit mathematically identical formulations are provided. One of the formulations can be directly applied to design problems that arise in MIMO communication systems as will be demonstrated in Section 10.4.

In short, the contributions of this chapter are two-fold. Firstly, we show that the solution is independent from the shape of the cost function $\mathcal{F}$; it is completely determined by the majorization constraint and the linear shift of the objective.

Secondly, we show that the problem is equivalent to calculating the convex hull of a simple polygon in $\mathbb{R}^2$, for which there exists efficient solvers with $O(N)$ complexity [MA79]. A solver tailored for this particular application is proposed.

A rigorous definition of the considered problem is presented in Section 10.2. In Section 10.3, we show that the problem is equivalent to the convex-hull problem of a simple polygon in $\mathbb{R}^2$. In Section 10.4, it is demonstrated how this problem can be applied to a MIMO communications problem, and finally the chapter is concluded in Section 10.5.

10.1.1 Notation

In addition to the notation introduced in Chapter 1, we will use the following notation: Given a vector $\boldsymbol{a}$, denote the prefix-sum vector as $\breve{\boldsymbol{a}}$ defined by $\breve{a}_k = \sum_{n=1}^{k} a_n$. For later notational simplicity we define $\breve{a}_0 = 0$. Denote by $\mathcal{D}$ the set of decreasing vectors, $\mathcal{D} = \left\{\boldsymbol{a} \in \mathbb{R}^N : a_1 \geq \ldots \geq a_N\right\}$, and the set of increasing vectors by $\mathcal{D}^* = \left\{\boldsymbol{a} \in \mathbb{R}^N : -\boldsymbol{a} \in \mathcal{D}\right\}$.

10.2 Problem formulation

We consider minimizing a cost function $\mathcal{F} : \mathcal{X} \mapsto \mathbb{R}$ with the property that it is Schur-convex on the symmetric domain[1] $\mathcal{X} \subset \mathbb{R}^N$ of interest, or equivalently,

$$\boldsymbol{z}_1 \preceq \boldsymbol{z}_2 \Longrightarrow \mathcal{F}(\boldsymbol{z}_1) \leq \mathcal{F}(\boldsymbol{z}_2), \quad \forall\, \boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{X}.$$

Herein, the optimization variable $\boldsymbol{z}$ is constrained to a convex domain defined by

$$\mathcal{Z} = \{\boldsymbol{z} : \boldsymbol{z} + \boldsymbol{w} \preceq \boldsymbol{c}\}, \tag{10.1}$$

where $\boldsymbol{w}$ and $\boldsymbol{c}$ are two vectors in $\mathbb{R}^N$. Due to the asymmetry that is introduced by the vector $\boldsymbol{w}$, we choose to refer to the constraint imposed by $\mathcal{Z}$ as a skewed majorization constraint. From here on it is assumed that $\boldsymbol{w}$ and $\boldsymbol{c}$ are chosen such that $\mathcal{Z} \subset \mathcal{X}$. Assume further that $\mathcal{F}$ is continuously differentiable on the interior of $\mathcal{Z}$. A few implications of these statements that will be useful in the following discussion are:

P-1. $\mathcal{F}$ is symmetric[2] on $\mathcal{X}$,

P-2. $\mathcal{F}$ is convex[3] on $\mathcal{X}$,

[1] A set $\mathcal{X}$ is symmetric if and only if $\boldsymbol{\Pi z} \in \mathcal{X}$ for all permutation matrices $\boldsymbol{\Pi}$ and all vectors $\boldsymbol{z} \in \mathcal{X}$.

[2] A function $\mathcal{F}$ is symmetric on $\mathcal{X}$ if for any $\boldsymbol{z} \in \mathcal{X}$ and any N-dimensional permutation matrix $\boldsymbol{\Pi}$, the function stays invariant to the permutation $\mathcal{F}(\boldsymbol{z}) = \mathcal{F}(\Pi \boldsymbol{z})$.

[3] A function $\mathcal{F}$ is convex on $\mathcal{X}$ if; for any two $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{X}$, and for all $\theta \in [0, 1]$, the following relation holds

$$\mathcal{F}\Big(\boldsymbol{z}_1\theta + \boldsymbol{z}_2(1-\theta)\Big) \leq \mathcal{F}(\boldsymbol{z}_1)\theta + \mathcal{F}(\boldsymbol{z}_2)(1-\theta).$$

P-3. $\boldsymbol{z} \in \mathcal{D}^* \cap \mathcal{X} \Longrightarrow \nabla\mathcal{F}(\boldsymbol{z}) \in \mathcal{D}^*$,

where Schur-convexity of $\mathcal{F}$ implies P-1 and P-2 [JB06, MO79], and also P-3 in conjunction with the additional assumption on differentiability.

The optimization problem of interest is given by

$$\min_{\boldsymbol{z}\in\mathcal{Z}} \mathcal{F}(\boldsymbol{z}), \tag{10.2}$$

with vectors $\boldsymbol{w}$ and $\boldsymbol{c}$ defining the domain according to (10.1). Without loss of generality we assume that $\boldsymbol{c}, \boldsymbol{w} \in \mathcal{D}$. This is true for $\boldsymbol{c}$ since the order of majorization is unaffected by permutations. For $\boldsymbol{w}$, the symmetry property, P-1, implies that Problem (10.2) with an unordered $\boldsymbol{w}$ can be solved via the equivalent problem with $\boldsymbol{w}$ permuted into decreasing order.

Problem (10.2) belongs to the class of convex optimization problems since the function $\mathcal{F}$ to be minimized is convex (P-2), and the domain $\mathcal{Z}$ is a convex set. The latter can be verified by equivalently writing the majorization constraint in (10.1) in terms of an affine equality constraint

$$\breve{z}_N + \breve{w}_N = \breve{c}_N, \tag{10.3}$$

together with a number of affine inequality constraints

$$\sum_{n\in\mathcal{I}} z_n + w_n \leq \breve{c}_k \quad \forall\, \mathcal{I} \subset \{1,\ldots,N\}\,, \tag{10.4}$$

in which k is the cardinality of $\mathcal{I}$. In this case, as the number of inequality constraints in (10.4) grows exponentially with N, the problem does not scale well with the number of dimensions. So, even though convex problems are sometimes conceived as easy to solve, this is an example of a fairly difficult convex problem.

10.2.1 The relaxed problem

We shall approach problem (10.2) by considering, for a while, a relaxation of the same problem by extending the feasible region. For each k in (10.4), we let $\mathcal{Z}_{\text{relax}}$ exclude all inequality constraints except those corresponding to $\mathcal{I} = \{1, 2, \ldots, k-1, k\}$. Introducing the vector $\boldsymbol{b} = \boldsymbol{c} - \boldsymbol{w}$, $\mathcal{Z}_{\text{relax}} \supset \mathcal{Z}$ is defined as

$$\mathcal{Z}_{\text{relax}} = \left\{ \boldsymbol{z} : \breve{\boldsymbol{z}} \leq \breve{\boldsymbol{b}},\ \breve{z}_N = \breve{b}_N \right\}. \tag{10.5}$$

Note that if the optimal solution $\boldsymbol{z}^*$ of the relaxed problem

$$\min_{\boldsymbol{z}\in\mathcal{Z}_{\text{relax}}} \mathcal{F}(\boldsymbol{z}), \tag{10.6}$$

also satisfies $\boldsymbol{z}^* \in \mathcal{Z}$, then $\boldsymbol{z}^*$ solves the original problem (10.2) as well.

10.3 Method to find the optimal point

The optimization problem (10.6) is convex (in fact so is (10.2)), hence the extensive framework of convex optimization, [BV04], can be applied to efficiently calculate the optimal solution numerically. In addition to this, for the case herein, we will show that the optimal solution decouples from the specific cost function that is used. Clearly, this will further simplify the optimization procedure, and an algorithm with $O(N)$ complexity is presented that determines the optimum exactly. We begin our analysis of the problem with the KKT optimality conditions.

10.3.1 The KKT optimality conditions

The Lagrangian cost function of the relaxed problem is

$$\mathcal{L}(\boldsymbol{z}, \boldsymbol{\lambda}) = \mathcal{F}(\boldsymbol{z}) + \boldsymbol{\lambda}^{\mathrm{T}} \left(\breve{\boldsymbol{z}} - \breve{\boldsymbol{b}} \right), \tag{10.7}$$

with Lagrangian multipliers $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_N]^{\mathrm{T}}$, which gives the KKT optimality conditions [BV04]:

$$\nabla \mathcal{F}(\boldsymbol{z})_n + \sum_{i=n}^{N} \lambda_i = 0, \quad 1 \leq n \leq N, \tag{10.8a}$$

$$\lambda_n \geq 0, \quad 1 \leq n \leq N-1, \tag{10.8b}$$

$$\breve{z}_n \leq \breve{b}_n, \quad 1 \leq n \leq N-1, \tag{10.8c}$$

$$\breve{z}_N = \breve{b}_N, \tag{10.8d}$$

$$\boldsymbol{\lambda}^{\mathrm{T}} \left(\breve{\boldsymbol{b}} - \breve{\boldsymbol{z}} \right) = 0. \tag{10.8e}$$

Slater's theorem (see e.g. [BV04]) states that if the problem is convex and there exists a strictly feasible point, $\boldsymbol{z}^f$, then strong duality holds and the KKT conditions are both sufficient and necessary for optimality. Such a point $\boldsymbol{z}^f$ can readily be found by defining α to be any constant such that $\alpha < \min_i b_i$. Then the vector $\boldsymbol{z}^f$, defined as

$$z_k^f = \begin{cases} \alpha & k = 1, \ldots, N-1, \\ (1-N)\alpha + \breve{b}_N & k = N, \end{cases} \tag{10.9}$$

satisfies $\breve{z}_k^f < \breve{b}_k$ for $k < N$ and $\breve{z}_N^f = \breve{b}_N$.

The KKT condition (10.8a) above depends on the cost function $\mathcal{F}(\boldsymbol{z})$ that is used. Next, we show that this dependency disappears when restricting $\mathcal{F}(\boldsymbol{z})$ to the class of Schur-convex objectives.

Theorem 1 *For a Schur-convex objective, $\mathcal{F}(\boldsymbol{z})$, the KKT conditions of the relaxed problem are fulfilled for a $\boldsymbol{z}$ satisfying the following constraints*

C-1. $\boldsymbol{z} \in \mathcal{D}^*$,

C-2. $\check{z}_n \leq \check{b}_n,\ 1 \leq n \leq N-1,$

C-3. $\check{z}_N = \check{b}_N,$

C-4. $z_n = z_{n+1}$ *if* $\check{z}_n < \check{b}_n,\ 1 \leq n \leq N-1.$

Proof: Conditions C-2 and C-3 are identical to the corresponding KKT conditions. KKT condition (10.8a) is satisfied by defining $\boldsymbol{\lambda}$ as

$$\lambda_k = \begin{cases} \nabla\mathcal{F}(\boldsymbol{z})_{k+1} - \nabla\mathcal{F}(\boldsymbol{z})_k & k = 1, \ldots, N-1, \\ -\nabla\mathcal{F}(\boldsymbol{z})_N & k = N. \end{cases} \tag{10.10}$$

Since $\mathcal{F}(\boldsymbol{z})$ is Schur-convex, condition C-1 implies that

$$\boldsymbol{z} \in \mathcal{D}^* \Longrightarrow \nabla\mathcal{F}(\boldsymbol{z}) \in \mathcal{D}^* \Longrightarrow \lambda_k \geq 0, \tag{10.11}$$

for all $k = 1, \ldots, N-1$ satisfying (10.8b). If $\check{z}_k < \check{b}_k$ for any $k = 1, \ldots, N-1$, then condition C-4 implies (10.8e) as

$$z_k = z_{k+1} \Longrightarrow \nabla\mathcal{F}(\boldsymbol{z})_k = \nabla\mathcal{F}(\boldsymbol{z})_{k+1} \Longrightarrow \lambda_k = 0. \tag{10.12}$$

□

The conditions in Theorem 1 are therefore sufficient to fulfill the KKT conditions. Consequently, due to Slater's theorem, any vector $\boldsymbol{z}$ that satisfies conditions C-1 to C-4 must be a global optimum of the relaxed problem (10.6). It remains to show that the conditions are also necessary, or in other words, that for any vector $\boldsymbol{b}$ there always exists a vector $\boldsymbol{z}$ that satisfies the constraints. This is shown constructively in Section 10.3.2.

The question now is whether the optimal solution, $\boldsymbol{z}$, of the relaxed problem (10.6) also satisfies the solution to the original problem (10.2). Recall that this is true if the optimal solution $\boldsymbol{z}$ satisfies $\boldsymbol{z} + \boldsymbol{w} \preceq \boldsymbol{c}$. This is implied by $\check{\boldsymbol{z}} + \check{\boldsymbol{w}} \leq \check{\boldsymbol{c}}$ provided that $\boldsymbol{z} + \boldsymbol{w} \in \mathcal{D}$, or more compactly, $\mathcal{Z} \supset \mathcal{Z}_{\text{relax}} \cap \mathcal{D}$. The following theorem reveals that the optimal $\boldsymbol{z}+\boldsymbol{w}$ is in fact a decreasing vector, provided that $\boldsymbol{w}$ and $\boldsymbol{c}$ are ordered in decreasing order.

Theorem 2 *If* $\boldsymbol{z}$ *satisfies the optimality conditions in Theorem 1 for* $\boldsymbol{b} = \boldsymbol{c} - \boldsymbol{w}$*, where* $\boldsymbol{c} \in \mathcal{D}$ *and* $\boldsymbol{w} \in \mathcal{D}$*, then* $\boldsymbol{z} + \boldsymbol{w} \in \mathcal{D}$*.*

Proof: Since the optimality conditions are satisfied, $\check{z}_n + \check{w}_n \leq \check{c}_n$ for all $1 \leq n \leq N-1$. Assume first that $\check{z}_n + \check{w}_n < \check{c}_n$. Then from C-4 we have

$$z_n + w_n = z_{n+1} + w_n \geq z_{n+1} + w_{n+1}. \tag{10.13}$$

If, on the other hand, $\check{z}_n + \check{w}_n = \check{c}_n$, then

$$\left.\begin{array}{rl} \check{z}_{n-1} + \check{w}_{n-1} & \leq \check{c}_{n-1} \\ \check{z}_n + \check{w}_n & = \check{c}_n \\ \check{z}_{n+1} + \check{w}_{n+1} & \leq \check{c}_{n+1} \end{array}\right\} \Longrightarrow$$

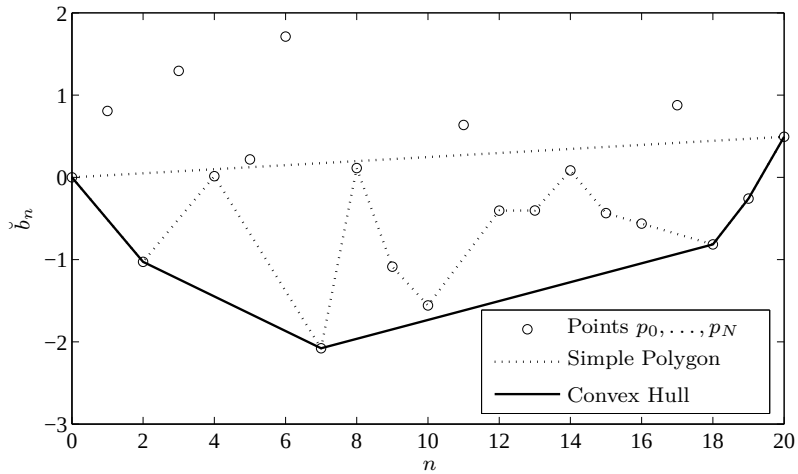

Figure 10.1: An illustration of a simple polygon of a sequence of points, and the convex hull under the same sequence.

$$\left.\begin{array}{rl} z_n + w_n & \geq c_n \\ z_{n+1} + w_{n+1} & \leq c_{n+1} \end{array}\right\} \Longrightarrow$$
$$z_n + w_n \geq z_{n+1} + w_{n+1}$$

where the last inequality uses the fact that $c_n \geq c_{n+1}$. □

10.3.2 Algorithms producing the optimal solution

In this section we prove that the problem of obtaining a $\boldsymbol{z}$ satisfying the modified KKT conditions can be solved by any algorithm identifying the convex hull of a finite set of points in $\mathbb{R}^2$. Moreover, we show that this can be performed with computational complexity of order $O(N)$. In the appendix, we present an easily implemented algorithm that avoids operating explicitly in $\mathbb{R}^2$.

In order to show that, in essence, solving the relaxed problem corresponds to a convex-hull problem, we refer to the framework used in [MA79]. A simple polygon is a sequence of points that, when connected with straight lines, partitions $\mathbb{R}^2$ into two disjoint sets. As guidance in the following discussion, Figure 10.1 illustrates the distinction between a sequence of points, a simple polygon, and the convex hull under the simple polygon respectively.

Theorem 3 *The $\boldsymbol{z}$ which solves* (10.2) *and* (10.6) *can be found using any algorithm that identifies the convex hull of a simple polygon in $\mathbb{R}^2$.*

Proof: By embedding the components of the vector $\breve{\boldsymbol{b}}$ into $\mathbb{R}^2$ according to

$$p_n = (n, \breve{b}_n) \quad \forall\, n = 0, \ldots, N,$$

it is obvious that either $p_0 \dots p_N$ is a simple polygon, or there are points above the line $p_0 p_N$ that may be removed to form a simple polygon.

A convex-hull algorithm produces a simple, convex polygon $p_{n_0} \dots p_{n_q}$, which we assume to be represented by a set of indices satisfying $n_r < n_{r+1}$, and necessarily, $n_0 = 0$ and $n_q = N$. The convexity property explicitly states that the slope between points is increasing $\forall r = 1, \dots, q-1$,

$$\frac{\breve{b}_{n_{r+1}} - \breve{b}_{n_r}}{n_{r+1} - n_r} > \frac{\breve{b}_{n_r} - \breve{b}_{n_{r-1}}}{n_r - n_{r-1}} \tag{10.14}$$

and the hull property that the straight line between neighboring points in the hull forms a lower bound, i.e. $\forall r = 0, \dots, q-1$,

$$\breve{b}_n \geq \breve{b}_{n_r} + (n - n_r)\frac{\breve{b}_{n_{r+1}} - \breve{b}_{n_r}}{n_{r+1} - n_r} \quad \forall n : n_r < n < n_{r+1}. \tag{10.15}$$

We are now ready to use the convex simple polygon, specified by the set of polygon indices $\mathcal{I}_A = \cup_r \{n_r\}$, to construct a vector $\boldsymbol{z}$ that satisfies the optimality conditions. Firstly, we enforce equality at the points of the polygon, i.e., $\breve{z}_n = \breve{b}_n$ if $n \in \mathcal{I}_A$. Then, in order to satisfy condition C-4 we make the slope of $\breve{z}_n$ constant for indices between polygon points,

$$z_n = \frac{\breve{b}_{n_{r+1}} - \breve{b}_{n_r}}{n_{r+1} - n_r}, \quad \text{if } n_r < n \leq n_{r+1}. \tag{10.16}$$

Clearly, the vector $\boldsymbol{z}$ is now completely determined from the corresponding polygon. Since condition C-4 is satisfied by definition, it remains to be shown that conditions C-1 to C-3 are satisfied as well. Again, by definition, C-3 is fulfilled since $N \in \mathcal{I}_A$. Pick an arbitrary index, $n \in \{1, \dots, N-1\}$, and identify the corresponding polygon point, r, satisfying $n_r < n \leq n_{r+1}$. If $n \in \mathcal{I}_A$ then $\breve{z}_n = \breve{b}_n$, while if $n \notin \mathcal{I}_A$ then by (10.15)

$$\breve{z}_n = \breve{b}_{n_r} + (n - n_r)\frac{\breve{b}_{n_{r+1}} - \breve{b}_{n_r}}{n_{r+1} - n_r} \leq \breve{b}_n. \tag{10.17}$$

Hence, C-2 is satisfied. Moreover, if $n \notin \mathcal{I}_A$ then $z_n = z_{n+1}$, while if $n \in \mathcal{I}_A$ then

$$z_n = \frac{\breve{b}_{n_r} - \breve{b}_{n_{r-1}}}{n_r - n_{r-1}} < \frac{\breve{b}_{n_{r+1}} - \breve{b}_{n_r}}{n_{r+1} - n_r} = z_{n+1}, \tag{10.18}$$

and we conclude that $\boldsymbol{z} \in \mathcal{D}^*$ (C-1). □

The general problem of identifying the convex hull of a set of points in $\mathbb{R}^2$ typically requires $O(N \log N)$ iterations. However, since the points are ordered as a simple polygon, the problem has additional structure that we can use. In [MA79], an algorithm was presented that solves the simple-polygon problem in $O(N)$ operations. For the interested reader, a simplified $O(N)$ algorithm is presented in Appendix 10.B, tailored for solving the particular problem considered herein.

10.4 Application to a communications problem

This section presents two applications to the optimization problem that is treated in this chapter. The problems concern transceiver design based on either perfect TX-CSI or partial TX-CSI of the correlation statistics of the channel. The receiver is assumed to know the channel perfectly and apply decision feedback detection. For the full details on optimal filter design for a DF MIMO system we refer to Chapter 8.

Consider the system model (2.1) in combination with linear precoding (2.25), where the transmitted signal is constructed using linear precoding of the data-signal vector $\boldsymbol{s} \in \mathbb{C}^{N_t}$ as

$$\boldsymbol{x} = \boldsymbol{F}\boldsymbol{s}.$$

The elements of $\boldsymbol{s}$ consist of uncorrelated modulated data signals of various constellation types, and the vector is normalized as

$$\mathrm{E}[\boldsymbol{s}\boldsymbol{s}^*] = \boldsymbol{I}.$$

Herein we restrict the precoder $\boldsymbol{F}$ to be scaled unitary, which — in combination with heterogenous signal constellations — is a reasonable restriction. Note that a unitary precoder corresponds to equal power allocation, which is close to optimal in the high-SNR region [YC01], as well as robust to errors in the channel estimate at the transmitter [BO06]. It also limits the maximum transmitted power per antenna which sometimes is restricted due to hardware constraints. We will now consider the two cases, perfect TX-CSI and partial TX-CSI, separately:

10.4.1 Perfect TX-CSI

The transmitter, as well as the receiver, is assumed to have perfect a-priori knowledge about the channel matrix $\boldsymbol{H}$. Define the matrix

$$\boldsymbol{Q} = (\boldsymbol{F}^*\boldsymbol{H}^*\boldsymbol{H}\boldsymbol{F} + \boldsymbol{I})^{-1}, \tag{10.19}$$

and denote the singular value decomposition as

$$\boldsymbol{Q} = \boldsymbol{U}\big((P/N_t)\boldsymbol{\Lambda}_{\boldsymbol{H}}^2 + \boldsymbol{I}\big)^{-1}\boldsymbol{U}^*. \tag{10.20}$$

Denote also the Cholesky decomposition as $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^*$, where $\boldsymbol{L}$ is a lower triangular matrix. The squared diagonal elements of $\boldsymbol{L}$ are of particular interest and are denoted the Cholesky elements. Define the vector

$$\boldsymbol{\xi} = \log\big(|\boldsymbol{d}(\boldsymbol{L})|^{-2}\big). \tag{10.21}$$

By assuming the filters in the DF detector are designed to minimize the MSE, the vector of MSEs prior symbol detection is [PJ07, Sec. 4.3]

$$\boldsymbol{\alpha} = \exp(-\boldsymbol{\xi}). \tag{10.22}$$

The spatial subchannels are not equally sensitive to noise if different types of signal constellations are used. This fact is important to take into account in the precoder design, and one way to do this is to form an objective function consisting of weighted MSEs. Typically we choose the weights to be proportional to the inverse of the squared minimum distance of the constellation used. Define the vector of the logarithm of these weights as $\boldsymbol{w}$. Suppose we would like to minimize a cost function based on the weighted MSE, our objective function is

$$\mathcal{F}\big(\exp(\boldsymbol{w}-\boldsymbol{\xi})\big), \tag{10.23}$$

where $\mathcal{F}\big(\exp(\cdot)\big)$ is a Schur-convex function. At first glance the problem may look difficult due to the unknown relation between the eigenvectors, $\boldsymbol{U}$, of the matrix $\boldsymbol{Q}$, and the Cholesky elements. However, it is known that the logarithm of the Cholesky elements are always majorized by the logarithm of the singular values [HJ91]. Consequently, if and only if the Cholesky elements satisfies

$$\boldsymbol{\xi} \preceq \log \boldsymbol{d}\big((P/N_t)\boldsymbol{\Lambda}_{\boldsymbol{H}}^2+\boldsymbol{I}\big) \triangleq \boldsymbol{c}, \tag{10.24}$$

where $\preceq$ denotes majorization, then there exists a unitary matrix $\boldsymbol{U}$, and a Cholesky factor $\boldsymbol{L}$, such that $\boldsymbol{U}\big((P/N_t)\boldsymbol{\Lambda}_{\boldsymbol{H}}^2+\boldsymbol{I}\big)^{-1}\boldsymbol{U}^*=\boldsymbol{L}\boldsymbol{L}^*$. The problem of designing the precoder, $\boldsymbol{F}$, can therefore be reformulated to the form

$$\begin{array}{ll}\underset{\boldsymbol{\xi}}{\text{minimize}} & \mathcal{F}\big(\exp(\boldsymbol{w}-\boldsymbol{\xi})\big) \\ \text{subject to} & \boldsymbol{\xi} \preceq \boldsymbol{c}\end{array}. \tag{10.25}$$

As is shown in Appendix 10.A, problem (10.32) is equivalent to the original problem formulation (10.2). Once the optimal $\boldsymbol{\xi}$ has been calculated (e.g., using the algorithm in Appendix 10.B), the unitary matrix $\boldsymbol{U}$ can be calculated using the GTD [PJ07, Appendix B].

10.4.2 Partial TX-CSI

In this case the transmitter has access to first and second order statistics of the channel. More specifically the channel matrix — as seen from the transmitter — is assumed to be drawn from the following distribution

$$\text{vec}(\boldsymbol{H}) \sim CN(\boldsymbol{0}, \boldsymbol{R}_t^{\text{T}} \otimes \boldsymbol{I}), \tag{10.26}$$

where $\boldsymbol{R}_t$ is the correlation from the transmitter side, i.e. $\boldsymbol{R}_t = \text{E}[\boldsymbol{H}^*\boldsymbol{H}]$. Using a zero forcing DF receiver, it was shown in [LZW08] that the MSEs can be computed from the Cholesky elements of $\boldsymbol{L}\boldsymbol{L}^* = \boldsymbol{F}^*\boldsymbol{R}_t\boldsymbol{F}$ as

$$\alpha_i = \exp\big(-\xi_i - \log(N_r - i)\big) \quad \forall\, i = 1, ..., N_t, \tag{10.27}$$

where

$$\boldsymbol{\xi} \triangleq \log\big(|\boldsymbol{d}(\boldsymbol{L})|^2\big). \tag{10.28}$$

Note that the weights in (10.27) require us to use more receive antennas than transmit antennas, such that $N_r > N_t$. Again, if we use different signal constellations on the subchannels it is a good idea to weight the MSEs depending on the sensitivity of the constellations. Denote these weights $\boldsymbol{w}$, and define a cost function as

$$\boldsymbol{\alpha}^{\mathrm{T}} \exp(\boldsymbol{w}). \tag{10.29}$$

Introduce the modified weights $\tilde{\boldsymbol{w}}$ as $\tilde{w}_i = w_i - \log(N_r - i)$ for all $i = 1, ..., N_t$. Then the objective function becomes

$$\mathbf{1}^{\mathrm{T}} \exp(\tilde{\boldsymbol{w}} - \boldsymbol{\xi}). \tag{10.30}$$

Assuming only scaled unitary precoders gives us the following sufficient and necessary condition for $\boldsymbol{\xi}$:

$$\boldsymbol{\xi} \preceq \log \boldsymbol{d}\big((P/N_t)\boldsymbol{\Lambda}_{\boldsymbol{R}_t}\big) \triangleq \tilde{\boldsymbol{c}}, \tag{10.31}$$

where $\boldsymbol{\Lambda}_{\boldsymbol{R}_t}$ is the diagonal matrix containing the singular values from $\boldsymbol{R}_t$. Similar to the previous case, the problem of designing the precoder, $\boldsymbol{F}$, can be reformulated to the form

$$\begin{array}{ll} \underset{\boldsymbol{\xi}}{\text{minimize}} & \mathbf{1}^{\mathrm{T}} \exp(\tilde{\boldsymbol{w}} - \boldsymbol{\xi}) \\ \text{subject to} & \boldsymbol{\xi} \preceq \tilde{\boldsymbol{c}} \end{array}. \tag{10.32}$$

Again, once the optimal $\boldsymbol{\xi}$ has been calculated, the unitary matrix $\boldsymbol{U}$ can be calculated using the GTD.

10.5 Conclusions

This chapter considered the problem of optimizing a Schur-convex objective under a skewed majorization constraint. It was shown that the solution does not depend on the shape of the cost function. But unlike the case with regular majorization constraints, for which the solution corresponds to having all elements equal, skewed constraints have greater impact on the optimal solution. It is also shown that the optimization problem is equivalent to identifying the convex hull under a simple polygon defined by the constraint parameters. This result allows us to calculate the solution instantaneously. As an application of the posed problem, a novel precoder design for MIMO communication systems utilizing decision feedback detection at the receiver was presented. Using the theoretical results herein, it was demonstrated how to efficiently calculate a precoder that take heterogenous signal constellations into account for the cases of perfect as well as partial TX-CSI.

Appendix 10.A Alternative problem formulations

By simple variable substitutions we can reformulate problem (10.2) as

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \mathcal{F}(\boldsymbol{x}-\boldsymbol{w}), \\ \text{st.} \quad & \boldsymbol{x} \preceq \boldsymbol{c}, \end{aligned} \tag{10.33}$$

where the cost function consists of a linearly shifted Schur-convex objective. Clearly, the objective is still convex but not symmetric due to the linear shift. Another form is

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \mathcal{G}(\alpha_1 x_1, \ldots, \alpha_N x_N), \\ \text{st.} \quad & \boldsymbol{x} \preceq_{\times} \boldsymbol{c}, \\ & \boldsymbol{x} \geq 0, \end{aligned} \tag{10.34}$$

where the α_i's are strictly positive weighting coefficients, and $\mathcal{F} = \mathcal{G} \circ \exp$ is Schur-convex (which is implied if $\mathcal{G}$ is Schur-convex).

Appendix 10.B Algorithm 2

Algorithm 2 determines the indices that correspond to points on the one-sided convex hull under $\boldsymbol{x}$. The stack vector of indices in the convex hull is denoted

Algorithm 2

1: $\boldsymbol{i}[0] \leftarrow 0$
2: $N \leftarrow \text{length}(\boldsymbol{x}) - 1$
3: $m \leftarrow 0$
4: **for** $j = 1$ **to** $N-1$ **do**
5: **if** $\frac{x_{\boldsymbol{i}[m]} - x_j}{j - \boldsymbol{i}[m]} > \frac{x_j - x_N}{N-j}$ **then**
6: **while** $m > 0$ **and** $\frac{x_{\boldsymbol{i}[m-1]} - x_{\boldsymbol{i}[m]}}{\boldsymbol{i}[m] - \boldsymbol{i}[m-1]} \leq \frac{x_{\boldsymbol{i}[m]} - x_j}{j - \boldsymbol{i}[m]}$ **do**
7: $m \leftarrow m - 1$
8: **end while**
9: $m \leftarrow m + 1$
10: $\boldsymbol{i}[m] \leftarrow j$
11: **end if**
12: **end for**
13: $m \leftarrow m + 1$
14: $\boldsymbol{i}[m] \leftarrow N$

$\boldsymbol{i}$. After the algorithm has terminated the stack contains $m + 1 \leq N$ elements. Line 1 initializes the stack with the first index 0 (remember, the first and the last points of $\boldsymbol{x}$ are always in the convex hull). In line 4, a loop over the elements of $\boldsymbol{x}$ is commenced. If the point at x_j is convex given that straight lines are drawn

between $x_{\boldsymbol{i}[m]}$, x_j, and x_N, then the index j must be appended to the stack vector of hull-indices, $\boldsymbol{i}$. Convexity is checked in the query on line 5. When appending a new index, we must verify whether any previous points in the hull under construction turn concave due to the new point x_j. The loop commenced at line 6 removes all such points from $\boldsymbol{i}$ if they exist. Finally on line 14, the last point in $\boldsymbol{x}$ is appended to the convex hull. Although the loop is nested, in total at most $2N$ iterations take place; In the first loop an element can be added to the index set, while for the inner loop, in each iteration an element is removed. The inner loop can therefore not have more iterations (in total) than the outer loop, hence the complexity is $O(N)$.

Chapter 11

Thesis conclusions

This thesis considers the problem of joint optimization of the bit loading and linear precoder for a MIMO communication system. The objective was to obtain the Pareto optimum in terms of minimum transmit power and error probability for the case of delay-limited transmission. Two types of detectors that differ in complexity and performance have been considered; the optimal ML detector, and the suboptimal but less complex DF detector.

The capacity-optimal rate allocation (bit loading) and linear precoding strategy is to orthogonalize the channel into parallel non-interfering subchannels. Herein, where delay-limited transmission is considered, it is shown that orthogonalization is generally suboptimal when using the ML detector. This was shown constructively with a transceiver design comprising of three steps; bit loading, optimization of the unitary transmit-directivity matrix, and lattice basis reduction of the effective channel matrix. By using certain upper and lower bounds on the performance, we showed that densely packed non-orthogonal lattices can provide a gain, the so-called packing gain. The drawback with these dense lattices is the high kissing number that comes with the packing gain. The kissing number has a negative impact on the performance, and it was shown that if the SNR is low, the negative impact due to the kissing number outweighs the positive impact of the packing gain. The results regarding the packing gain where confirmed numerically where gains of several dB were attained compared with the best possible orthogonal transmission. Another result from our numerical experiments was that blind transmission works surprisingly well (when using the ML detector), at least if the Rayleigh distributed elements of the channel matrix are uncorrelated. For moderately low SNR, blind transmission even outperformed the optimized orthogonal transmission. This raises the question whether TX-CSI is really useful for improving the system performance, or, if its usefulness is more a matter of reducing the complexity of the detection problem at the receiver?

The second part of the thesis focuses on the transceiver design when using the suboptimal DF detector. An algorithm with linear complexity that computes the

optimal transmit-power allocation given a prescribed bit loading was presented. Then we switched focus to the joint optimization of both bit loading and DF filters by relaxing the set of possible bit rates. Our main result shows that the jointly optimal bit loading and linear precoder orthogonalizes the channel. Interestingly, this result makes the DF part of the DF detector obsolete, and a linear detector is therefore sufficient for optimal performance. From this result, we note that the packing gain that was attainable using ML detection is not possible to realize when using a DF detector. Finally, it was shown that when all subchannels operate at a high bit rate, then all bit loadings perform almost equally well. Therefore, as long as the correct number of subchannels are active, optimal bit loading is not particularly important for close-to-optimal performance when using a DF detector. Although, by using a suboptimal bit loading the feedback part of the DF receiver will clearly no longer be zero.

While working on optimization of DF filters we identified a certain class of optimization problems that can be solved very efficiently. The class of problems concerns optimization of a Schur-convex objective under a linearly shifted, or skewed, majorization constraint. Although the problem as it stands is convex, it does not scale well since the number of linear constraints grows exponentially with the number of dimensions. We showed that a simpler relaxed problem with the same optimum as the original problem can be considered instead. The solution to the problem was shown to be independent of the shape of the cost function, i.e. the solution is the same for the entire class of cost functions. Unlike the optimization problems with regular majorization constraints, for which the optimum corresponds to having all elements equal, skewed constraints impact the optimal solution more directly. The optimization problem was shown to be equivalent to the problem of identifying the convex hull under a simple polygon that is defined by the constraint parameters. This property of the optimal solution makes the problem particularly easy to solve. We presented two practical applications that arise in the field of MIMO communication systems: Using skewed majorization it was shown how to efficiently calculate a unitary precoder that takes heterogenous signal constellations into account for the cases of perfect as well as partial TX-CSI.

11.1 Future work

There are a number of questions related to the work herein that remain open. In the following we list a few of these issues.

- Problems regarding transceiver design for the ML detector are particularly difficult to solve. Nevertheless, the ML decoder is the optimal detector and improving the transceiver design that was presented herein is an interesting line for future research. Specifically, one may consider modifying the basis reduction algorithm (the LLL algorithm) to jointly optimize the bit loading together with the lattice basis vectors.

- It would be interesting to gain more insight into why blind transmission shows such promising performance when using the ML detector. The fact that blind transmission outperforms the optimal orthogonal transceiver for moderately low SNRs is somewhat surprising to us.
- The transceiver design problem based on partial or imperfect TX-CSI remains an open problem for both ML and the DF receiver. When the TX-CSI is no longer perfect, cross talk between the subchannels is inevitable, and DF will consequently outperform the linear detector. For the ML decoder our lattice based approach faces many difficulties when the channel matrix is not deterministic, for example a lattice with a high kissing number is likely to be more sensitive to perturbations than a lattice with a low kissing number. However, the good performance shown by the of blind transmission suggests that there should exist simple schemes that perform well even for the case of partial TX-CSI.
- The use of delay-limited transmission as opposed to capacity-optimal transmission needs more investigation. To do this, higher layers (e.g. the application layer) have to be considered. It would be interesting to compare various transmission schemes using different types of channel coding (or no channel coding) for a delay-sensitive application, such as the closed-loop control application.

Bibliography

[BB06] A. Bennatan and D. Burshtein. Design and analysis of nonbinary LDPC codes for arbitrary discrete-memoryless channels. *IEEE Transactions on Information Theory*, 52(2):549–583, February 2006.

[Bel73] E. Beltrami. Sulle funzioni bilineari. *Giornale De Matematiche ad Uso degli Studenti Delle Universita*, 11:98–106, 1873.

[BJJO08] S. Bergman, S. Järmyr, E. Jorswieck, and B. Ottersten. Optimization with skewed majorization constraints: Application to MIMO systems. In *Personal, Indoor and Mobile Radio Communications (PIMRC). IEEE International Symposium on*, pages 1–6, September 2008.

[BMO04] S. Bergman, C. Martin, and B. Ottersten. Bit and power loading for spatial multiplexing using partial channel state information. In *Proceedings ITG Workshop on Smart Antennas, Technische Universität Munich*, pages 152–159, March 2004.

[BO05a] S. Bergman and B. Ottersten. Adaptive spatial bit loading using imperfect channel state information. In *Proceedings of International Workshop on Optical and Electronic Device Technology for Access Networks, Aalborg, Denmark*, September 2005. Invited Paper.

[BO05b] S. Bergman and B. Ottersten. Spatial multiplexing over Rician fading channels: Linear precoding transmission strategies. In *Nordic Conference on Radio Science and Communications (RVK)*, June 2005.

[BO06] S. Bergman and B. Ottersten. Design of robust linear dispersion codes based on imperfect CSI for ML receivers. In *Proceedings European Signal Processing Conference*, September 2006.

[BO07] S. Bergman and B. Ottersten. Lattice based linear precoding for MIMO block codes. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages III329–III332, April 2007.

[BO08] S. Bergman and B. Ottersten. Lattice based linear precoding for multi-carrier block codes. *IEEE Transactions on Signal Processing*, 56(7):2902–2914, July 2008.

[BP79] C. A. Belfiore and J. H. Jr. Park. Decision feedback equalization. *Proceedings of the IEEE*, 67(8):1143–1156, August 1979.

[BPO08] S. Bergman, D. P. Palomar, and B. Ottersten. Joint bit allocation and precoding for MIMO systems with decision feedback detection. *IEEE Transactions on Signal Processing*, November 2008. Submitted to.

[BPO09] S. Bergman, D. P. Palomar, and B. Ottersten. Optimal bit loading for MIMO systems with decision feedback detection. In *Proceedings IEEE Vehicular Technology Conference*, April 2009. Invited paper.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[Cam99] J. Campello. Practical bit loading for DMT. In *IEEE Internationell Conference on Communications*, volume 2, pages 801–805, 1999.

[CBRB04] L. Collin, O. Berder, P. Rostaing, and G. Burel. Optimal minimum distance-based precoder for MIMO spatial multiplexing systems. *IEEE Transactions on Signal Processing*, 52(3):617–627, March 2004.

[CCB95] P. S. Chow, J. M. Cioffi, and J. A. C. Bingham. A practical discrete multitone transceiver loading algorithm for data transmission over spectrally shaped channels. *IEEE Transactions on Communications*, 43(234):773–775, February 1995.

[CDEF95] J. M. Cioffi, G. P. Dudevoir, M. V. Eyuboglu, and G. D. Forney. MMSE decision-feedback equalizers and coding—Part II: Coding results. *IEEE Transactions on Communications*, 43(10):2595–2604, October 1995.

[CS88] J. H. Conway and N. J. A. Sloane. *Sphere Packings, Lattices and Groups*. Springer-Verlag, 1988.

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[DDLW03] Y. Ding, T. N. Davidson, Z.-Q. Luo, and K. M. Wong. Minimum BER block precoders for zero-forcing equalization. *IEEE Transactions on Signal Processing*, 51(9):2410–2423, September 2003.

[DDW03] Y. Ding, T. N. Davidson, and K. M. Wong. On improving the BER performance of rate-adaptive block transceivers, with applications to DMT. In *Proceedings IEEE Global Telecommunications Conference*, volume 3, pages 1654–1658, December 2003.

[DGC03] M. O. Damen, H. E. Gamal, and G. Caire. On maximum-likelihood detection and the search for the closest lattice point. *IEEE Transactions on Information Theory*, 49(10):2389–2402, October 2003.

[DPSB07] E. Dahlman, S. Parkvall, J. Sköld, and P. Beming. *3G Evolution HSPA and LTE for Mobile Broadband*. Elsevier, 2007.

[FE91] G. D. Jr. Forney and M. V. Eyuboglu. Combined equalization and coding using precoding. *Communications Magazine IEEE*, 29(12):25–34, December 1991.

[FG98] G. J. Foschini and M. J. Gans. On limits of wireless communications in a fading environment when using multiple antennas. *Wireless Personal Communications*, 6(3):311–335, March 1998.

[FH96] R. F. H. Fischer and J. B. Huber. A new loading algorithm for discrete multitone transmission. In *Proceedings IEEE Global Telecommunications Conference*, volume 1, pages 724–728, November 1996.

[For99] G. David Forney. Multidimensional constellations—Part I: Introduction, figures of merit, and generalized cross constellations. *IEEE Journal on Selected Areas in Communications*, 7(6):877–892, August 1999.

[FWLH02a] R. F. H. Fischer, C. Windpassinger, A. Lampe, and J. B. Huber. MIMO precoding for decentralized receivers. In *Proc. IEEE Int. Symp. on Inf. Theory*, page 496, December 2002.

[FWLH02b] R. F. H. Fischer, C. Windpassinger, A. Lampe, and J. B. Huber. Space-time transmission using Tomlinson-Harashima precoding. In *Proceedings ITG Workshop on Smart Antennas, Berlin*, pages 139–147, January 2002.

[Gal62] R. Gallager. Low-density parity-check codes. *Information Theory, IRE Transactions on*, 8(1):21–28, January 1962.

[GC97] A. J. Goldsmith and Soon-Ghee Chua. Variable-rate variable-power MQAM for fading channels. *IEEE Transactions on Communications*, 45(10):1218–1230, October 1997.

[GC01] G. Ginis and J. M. Cioffi. On the relation between V-BLAST and the GDFE. *Communications Letters, IEEE*, 5(9):364–366, September 2001.

[Giv58] W. Givens. Computation of plane unitary rotations transforming a general matrix to triangular form. *Journal of SIAM*, 6(1):26–50, March 1958.

[Gra53] F. Gray. Pulse code communication. United States Patent, 2,632,058, March 1953.

[Gue03] T. Guess. Optimal sequences for CDMA with decision-feedback receivers. *IEEE Transactions on Information Theory*, 49(4):886–900, April 2003.

[Had93] J. Hadamard. Résolution d'une question relative aux déterminants. *Bulletin des Sciences Mathématiques*, 17:240–246, 1893.

[HH] D. Hughes-Hartogs. Ensemble modem structure for imperfect transmission media. United States Patents, 4,679,227 (July 1987), 4,731,816 (March 1998), 4,833,706 (May 1998).

[HH02] B. Hassibi and B. M. Hochwald. High-rate codes that are linear in space and time. *IEEE Transactions on Information Theory*, 48(7):1804–1824, July 2002.

[HJ85] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[HJ91] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

[HM72] H. Harashima and H. Miyakawa. Matched-transmission technique for channels with intersymbol interference. *IEEE Transactions on Communications*, 20(4):774–780, August 1972.

[Hou58] A. Householder. Unitary triangularization of a nonsymmetric matrix. *Journal of ACM 6*, 6:339–342, October 1958.

[HPS05] B. M. Hochwald, C. B. Peel, and A. L. Swindlehurst. A vector-perturbation technique for near-capacity multi-antenna multi-user communication–Part II: Perturbation. *IEEE Transactions on Communications*, 53(3):537–544, March 2005.

[JB06] E. Jorswieck and H. Boche. Majorization and matrix-monotone functions in wireless communications. *Foundations and Trends in Communications and Information Theory*, 3:553–701, 2006.

[JBO08] S. Järmyr, S. Bergman, and B. Ottersten. Long-term adaptive precoding for decision feedback equalization. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2897–2900, April 2008.

[JHL05] Y. Jiang, W. Hager, and J. Li. The geometric mean decomposition. *Linear Algebra and Its Applications*, 396:373–384, February 2005.

[JHL06] Y. Jiang, W. W. Hager, and J. Li. Tunable channel decomposition for MIMO communications using channel state information. *IEEE Transactions on Signal Processing*, 54(11):4405–4418, November 2006.

[JHL08] Y. Jiang, W. W. Hager, and J. Li. The generalized triangular decomposition. *Mathematics of Computation*, 77(262):1037–1056, April 2008.

[JO05] J. Jaldén and B. Ottersten. On the complexity of sphere decoding in digital communications. *IEEE Transactions on Signal Processing*, 53:1474–1484, April 2005.

[Jor74] C. Jordan. Mémoire sur les formes bilinéaires. *Journal de Mathématiques Pures et Appliquées, Deuxième Série*, 19:35–54, 1874.

[JSBC04] E. A. Jorswieck, A. Sezgin, H. Boche, and E. Costa. Optimal transmit strategies in MIMO Ricean channels with MMSE receiver. In *Vehicular Technology Conference*, volume 5, pages 3787–3791, September 2004.

[JSM08] J. Jalden, D. Seethaler, and G. Matz. Worst- and average-case complexity of LLL lattice reduction in MIMO wireless systems. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2685–2688, April 2008.

[KRJ98] B. S. Krongold, K. Ramchandran, and D. L. Jones. Computationally efficient optimal power allocation algorithm for multicarrier communication systems. In *IEEE Internationell Conference on Communications*, volume 2, pages 1018–1022, 1998.

[KS00] P. Kosowski and A. Smoktunowicz. On constructing unit triangular matrices with prescribed singular values. *Mathematics and Statistics*, 64(3):279–285, May 2000.

[KS04] M. Kiessling and J. Speidel. Statistical prefilter design for MIMO ZF and MMSE receivers based on majorization theory. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 313–316, May 2004.

[LLL82] A. K. Lenstra, H. W. Lenstra, and L. Lovasz. Factoring polynomials with rational coefficients. *IEEE Transactions on Information Theory*, 261(4):515–534, December 1982.

[LS03] E. G. Larsson and P. Stoica. *Space-time block coding for wireless communications*. Cambridge University Press, 2003.

[LYW04] B. Lu, G. Yue, and X. Wang. Performance analysis and design optimization of LDPC-coded MIMO OFDM systems. *IEEE Transactions on Signal Processing*, 52(2):348–361, February 2004.

[LZW08] T. Liu, J. Zhang, and K. M. Wong. Optimal precoder design for correlated MIMO systems using decision feedback receivers. In *Proceedings IEEE International Symposium on Information Theory*, pages 574–578, July 2008.

[MA79] D. McCallum and D. Avis. A linear algorithm for finding the convex hull of a simple polygon. *Information Processing Letters*, 9(5):201–206, December 1979.

[MBO04a] C. Martin, S. Bergman, and B. Ottersten. Simple spatial multiplexing based on imperfect channel estimates. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 713–716, May 2004.

[MBO04b] C. Martin, S. Bergman, and B. Ottersten. Spatial loading based on channel covariance feedback and channel estimates. In *Proceedings European Signal Processing Conference*, pages 519–522, September 2004.

[MO79] A. Marshall and I. Olkin. *Inequalities: Theory of Majorization and Its Applications*. Academic Press, Inc., 1979.

[MO04] C. Martin and B. Ottersten. Asymptotic eigenvalue distributions and capacity for MIMO channels under correlated fading. *IEEE Transactions on Wireless Communications*, 3(4):1350–1359, July 2004.

[PB05] D. P. Palomar and S. Barbarossa. Designing MIMO communication systems: Constellation choice and linear transciever design. *IEEE Transactions on Signal Processing*, 53(10):3804–3818, October 2005.

[PBO05] D. P. Palomar, M. Bengtsson, and B. Ottersten. Minimum BER linear transceivers for MIMO channels via primal decomposition. *IEEE Transactions on Signal Processing*, 53(8):2866–2882, August 2005.

[PCL03] D. P. Palomar, J. M. Cioffi, and M. A. Lagunas. Joint Tx-Rx beamforming design for multicarrier MIMO channels: A unified framework for convex optimization. *IEEE Transactions on Signal Processing*, 51(9):2381–2401, September 2003.

[PJ07] D. P. Palomar and Y. Jiang. *MIMO Tranceiver Design via Majorization Theory*. Now Publishers Inc., 2007.

[PLC04] D. P. Palomar, M. A. Lagunas, and J. M. Cioffi. Optimum linear joint transmit-receive processing for MIMO channels with QoS constraints. *IEEE Transactions on Signal Processing*, 52(5):1179–1197, May 2004.

[Pro01] J. G. Proakis. *Digital Communications*. McGraw-Hill, 2001.

[RU08] T. Richardson and R. Urbanke. *Modern Coding Theory.* Cambridge University Press, 2008.

[SA00] M. K. Simon and M.-S. Alouini. *Digital Communications Over Fading Channels.* New York: Wiley, 2000.

[SD07] M. B. Shenouda and T. N. Davidson. Minimum SER zero-forcing transmitter design for MIMO channels with interference pre-subtraction. pages 2109–2113, April 2007.

[SD08] M. B. Shenouda and T. N. Davidson. A framework for designing MIMO systems with decision feedback equalization or Tomlinson-Harashima precoding. *IEEE Journal on Selected Areas in Communications*, 26(2):401–411, February 2008.

[SFGK00] D.-S. Shiu, G. J. Foschini, M. J. Gans, and J. M. Kahn. Fading correlation and its effect on the capacity of multielement antenna systems. *Communications, IEEE Transactions on*, 48(3):502–513, March 2000.

[SGS01] A. Stamoulis, G. B. Giannakis, and A. Scaglione. Block FIR decision-feedback equalizers for filterbank precoded transmissions with blind channel estimation capabilities. *IEEE Transactions on Communications*, 49(1):69–83, January 2001.

[Sha48] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 623–656, 1948.

[tBKA04] S. ten Brink, G. Kramer, and A. Ashikhmin. Design of low-density parity-check codes for modulation and detection. *IEEE Transactions on Communications*, 52(4):670–678, April 2004.

[Tel95] E. Telatar. Capacity of multi-antenna Gaussian channels. *Technical Memorandum, Bell Laboratories (Published in European Transactions on Telecommunications, Vol. 10, No.6, pp. 585-595, Nov/Dec 1999)*, 1995.

[Tom71] M. Tomlinson. New automatic equaliser employing modulo arithmetic. *Electron. Lett.*, pages 138–139, March 1971.

[TSC98] V. Tarokh, N. Seshadri, and A. R. Calderbank. Space-time codes for high data rate wireless communication: performance criterion and code construction. *IEEE Transactions on Information Theory*, 44:744–765, March 1998.

[WF03] C. Windpassinger and Robert F. H. Fischer. Optimum and sub-optimum lattice-reduction-aided detection and precoding for MIMO communications. In *In Proceedings of the Canadian Workshop on Information Theory, Waterloo, Canada*, pages 88–91, May 2003.

[WFGV98] P. W. Wolniansky, G. J. Foschini, G. D. Golden, and R. A. Valenzuela. V-BLAST: An architecture for realizing very high data rates over the rich-scattering wireless channel. *URSI International Symposium on Signals, Systems, and Electronics*, pages 295–300, October 1998.

[XDZW06] F. Xu, T. N. Davidson, J.-K. Zhang, and K. M. Wong. Design of block transceivers with decision feedback detection. *IEEE Transactions on Signal Processing*, 54(3):965–978, March 2006.

[YBO+01] K. Yu, M. Bengtsson, B. Ottersten, D. McNamara, P. Karlsson, and M. Beach. Second order statistics of NLOS indoor MIMO channels based on 5.2 GHz measurements. In *Proceedings IEEE Global Telecommunications Conference*, volume 1, pages 156–160, 2001.

[YC01] W. Yu and J. M. Cioffi. On constant power water-filling. In *IEEE Internationell Conference on Communications*, volume 6, pages 1665–1669, 2001.

[ZT03] L. Zheng and D. N. C. Tse. Diversity and multiplexing: A fundamental tradeoff in multiple-antenna channels. *IEEE Transactions on Information Theory*, 49(5):1073–1096, May 2003.