MIMO Transceiver Design for Multi-Antenna Communications over Fading Channels

SIMON JÄRMYR

Doctoral Thesis in Telecommunications
Stockholm, Sweden 2013
Akademisk avhandling som med tillstånd av Kungliga Tekniska högskolan
framlägges till offentlig granskning för avläggande av teknologie doktors-
examen i telekommunikation måndagen den 22 april 2013 klockan 10.15 i
hörsal E2, Lindstedtsvägen 3, Stockholm.

© Simon Järmyr, April 2013, except where otherwise stated.

Many of the results have previously been published under IEEE copyright.

Tryck: Universitetsservice US-AB
Till Lovis...
Abstract

In wireless communications, the use of multiple antennas for both transmission and reception is associated with performance gains of fundamental nature. One such gain stems from the spatial-multiplexing capabilities of wireless multiple-input multiple-output (MIMO) channels: Many propagation environments admit several data streams to be conveyed in parallel over a single point-to-point link, setting the stage for significantly increased data rates. The optimal maximum-likelihood (ML) receiver pays a high price for spatial multiplexing in the form of a heavy computational burden. Another receiver candidate is the decision-feedback (DF) equalizer, reaping the benefits of spatial multiplexing with far more efficient receive processing. Its combination of independent data-stream decoding with successive interference cancellation (SIC) is sufficient to achieve several information-theoretic performance limits, both for point-to-point and multiple-access channels. Nevertheless, in practical systems there is often a clear performance degradation associated with DF processing compared to ML. In this context, linear precoding is a transmit pre-processing technique that can enhance performance by adapting the transmission to available channel-state information (CSI).

This thesis deals with MIMO transceiver design for DF equalization: the joint optimization of transmit precoding and receive equalization. A crucial aspect is the assumptions made on the kind of CSI available. The thesis considers a practical case with long-term, statistical CSI at the transmitter and perfect short-term CSI at the receiver. The thesis presents an optimization framework for MIMO transceiver design, contributing in several respects. Firstly, a number of relevant performance measures are presented, and novel expressions are provided for spatially correlated MIMO channels. Secondly, it is shown that optimization with respect to such performance measures can be cast as convex optimization problems, enabling efficient solutions in general. The thesis also provides an in-depth analysis of specific optimization problems, enabling very efficient solutions; novel connections are established between MIMO transceiver design, convex-hull algorithms, and submodular optimization. Thirdly, an extension to the multi-user uplink is provided. The thesis considers not only the joint optimization of the users’ precoders, but also joint optimization with the decoding order employed at the receiver. The thesis shows how to address the hard design problem in a computationally efficient manner using alternating optimization between precoders and the decoding order, which is observed to converge fast with close to optimal performance.
Acknowledgements

I am grateful to my supervisor Prof. Björn Ottersten for taking me on as a Ph.D. student in the Signal Processing group. He knows how excellent research is conducted, and has always encouraged me, watched over my progress, and guided me with gentle means. The freedom he has given me—to explore the problems that caught my curiosity—has made research a daily, exciting quest for answers.

I’m also grateful to Prof. Eduard Jorswieck for his solid support over the years as a co-author, always providing me with excellent feedback and insights. He was also the one that introduced me to the beautiful theory of majorization—an indispensable mathematical toolbox for solving many of the problems in this thesis. I’m also indebted to Prof. Mats Bengtsson, my co-advisor, for assisting me in so many ways, ranging from explaining obscure details of optimization theory to doing magical tricks in \LaTeX. I would also like to thank Dr. Svante Bergman for sharing his skills in the art of MIMO transceiver design, leading to exciting collaboration. I’ve always enjoyed the frequent discussions with Dr. Emil Björnson, my wing-mate during years of office sharing. He draws such lovely golden stars when reviewing my work. And I thank Alla Tarighati—my new office mate—for bearing with me during this last intense phase, and for always making me lighten up. I also thank Annika Augustsson and Tove Schwarz for taking such good care of the administration. And I thank all of you at the Signal Processing and Communication Theory groups who have contributed to the warm, and stimulating working atmosphere that I have highly appreciated.

I wish to acknowledge those who have helped me with proof-reading parts of the thesis: Rasmus Brandt, Jinghong Yang, Efthymios Tsakonas, Emil Björnson, David Hammarwall, and Bengt Samuelsson. I’m also grateful to Prof. Antonio Pascual-Iserte for the time and effort spent on acting as opponent, and to the grading committee: Prof. Kimmo Kansanen, Prof. Michail Matthaiou, and Prof. Tobias Oechtering.

Last, but certainly not least, I’d like to mention family and friends that mean so much to me (perhaps more than you know). In particular, my loving and ever-supporting parents Ingrid and Bengt have always been there for me (even proof-reading every single page of the thesis). And my two brothers Joel and David have always inspired me in various ways. I’m glad for my precious son Melker, who has an amazing ability to make me put away what seems so important, and do what’s really important: to just come and play. And finally, I thank my wife Lina for genuinely being you all along, the one I’ve always loved.

Simon Järmyr
Stockholm, March 2013
# Contents

1 Introduction 1  
1.1 The Wireless Channel 1  
1.2 Communication over Wireless Channels 3  
1.3 Designing MIMO Communication Systems 7  
1.4 Aim of the Thesis 10  

2 Problem Formulation and Contributions 11  
2.1 The Wireless MIMO System Model 11  
2.2 MIMO Transceiver Design 16  
2.3 Outline of Thesis Contributions 27  

3 A Framework for Precoder Optimization 31  
3.1 System Model 31  
3.2 Performance Measures 34  
3.3 Precoder Optimization 43  
3.4 Numerical Results 52  
3.5 Summary 57  
3.A A Collection of Proofs 58  

4 Performance Measures in Correlated MIMO Channels 65  
4.1 SNR Characterization 65  
4.2 Performance Under the Generalized Gamma Distribution 67  
4.3 SNR Approximations for ZF Receivers 74  
4.4 Numerical Results 77  
4.5 Summary 79  
4.A A Collection of Proofs 81  

5 Optimization with Majorization Constraints 87  
5.1 Optimization with Skewed Majorization Constraints 87  
5.2 Optimization with Submodular Constraints 94  
5.3 Statistical Precoding for ZF-DF Equalization 98  
5.4 Summary 102  
5.A A Convex-Hull Algorithm 104
6 Precoding and Ordering in the Multi-User Uplink 109
   6.1 System Model ............................................. 109
   6.2 User Utility Functions ................................. 114
   6.3 Utility Regions ........................................... 116
   6.4 Maximizing System Utility ............................... 120
   6.5 Efficient Single-Mode Selection ......................... 123
   6.6 Numerical Results ......................................... 130
   6.7 Summary .................................................... 136
   6.A A Collection of Proofs ................................. 137

7 Conclusions and Future Work 141
   7.1 Conclusions ............................................... 141
   7.2 Future Work ............................................... 142

Nomenclature 145

Bibliography 149
Chapter 1

Introduction

This chapter introduces basic elements of wireless communications, outlining the technical context of the thesis. In particular, it serves to illuminate the three components comprising the title of the thesis: fading channels, multi-antenna communications, and MIMO transceiver design. A more precise problem formulation is provided in the next chapter, along with an overview of the thesis’ contributions.

1.1 The Wireless Channel

Antennas are key elements for both transmitting and receiving radio waves that are commonly used for wireless communications. A transmit antenna converts a fast-oscillating electric current—representing a message to be sent—into radio waves that can move freely through air without being absorbed. Traveling at the speed of light, these waves may rapidly carry the message to a distant location. A tiny fraction of the emitted energy will find its way to the receiving antenna, exciting an electric current from which the message can be extracted.

The wireless channel describes the physical environment from the perspective of radio-wave propagation. Radio waves interact with objects and obstructions that are larger than or similar in size to the wavelength, which is typically on the order of decimeters or centimeters for cellular networks or WLAN. In an indoor environment, for example, there are walls, doorways, furniture, etc., determining how radio waves propagate. Three important principles of interactions between radio waves and surrounding objects are illustrated in Figure 1.1. Reflection occurs when radio waves impinge on a large, smooth surface such as a wall. Scattering may occur if the surface is very rough, spreading the reflected wave in multiple directions. Diffraction is a characteristic wave phenomenon, causing the impinging wave to bend around sharp edges.
CHAPTER 1. INTRODUCTION

1.1.1 Multipath Propagation

Objects interacting with radio waves give rise to multipath propagation; there may be many physical paths connecting two antennas, accounting for all reflections, scattering, and diffraction caused by the surroundings. Each path corresponds to a unique attenuation and phase-shift (or time-delay) of the transmitted radio wave. As illustrated in Figure 1.2, transmitted waves reach the receiver over different paths with different incident angles. When adding such waves, their relation in phase is crucial. Two waves “in phase” add constructively, while “out of phase” waves add destructively. The combined effect of multipath propagation may change substantially by moving the receiving device by half a wavelength due to changes in the phase associated with each path. Figure 1.3 illustrates an example of how the channel strength can vary due to multipath propagation. From the receiver’s point of view, the channel strength in its surroundings is a landscape of high hills and deep valleys.

The shape of the landscape depends on the frequency of the radio waves. This is a consequence of the fact that the different paths are of different length—the greater the difference in distance, the faster the change in frequency. It is common to speak of a coherence bandwidth: a range in frequency over which the landscape stays roughly the same.

1.1.2 Channel Fading

The variation in channel strength and phase over time is called channel fading. In the presence of multipath propagation, a receive antenna moving through the landscape illustrated in Figure 1.3 will experience the effect of multipath fading. Even if the receive antenna is fixed, similar fading due to multipath propagation will be present if the transmit antenna is moving, or even if only objects in the surroundings are moving.
1.2. COMMUNICATION OVER WIRELESS CHANNELS

Figure 1.2: In multipath propagation, transmitted waves reach the receiver from different directions with different phase shifts.

Figure 1.3: Illustration of spatial variations in channel strength due to multipath propagation.

Multipath fading is often referred to as small-scale fading, because this kind of channel variations belongs to the most rapid changes over time due to movement. If the transmitter or receiver travels a longer distance there will also be large-scale fading; that is, channel characteristics varying on a longer time scale. The contributions to such fading are commonly divided into two parts. The path-loss represents the average attenuation with the distance between the antennas, while shadowing accounts for additional attenuation due to large objects in the way.

1.2 Communication over Wireless Channels

With all kinds of wireless technology available today, the use of the frequency spectrum is strictly regulated. There are unlicensed frequency bands that are free to use, but the allowed transmit power is bounded and there may be substantial interference. This thesis focuses on licensed frequency bands, that are free from interference (and expensive to buy).

It is particularly simple to model the relation between the transmitted and received signals assuming narrowband communications; that is, when the available frequency bandwidth is less than the coherence bandwidth of the channel. The received signal is then comprised of two components—a desired signal term and an undesired noise term. The signal term is an attenuated and phase-shifted version of the transmitted signal as described in the previous section. The noise term accounts for the fact that all objects with a temperature emit radiation, causing an undesired contribution from
CHAPTER 1. INTRODUCTION

the surroundings to the received signal. The relative strength between the signal and noise terms, the signal-to-noise ratio (SNR), is a basic quality measure of the channel for the purpose of communication.

It was long believed that communicating messages over a noisy channel is inherently associated with a certain probability of making errors when recovering the messages. More reliable transmission can always be achieved by repetition coding, sending the same message over and over again. Unfortunately, this leads to a vanishing data rate as the number of repetitions increases. The picture changed with the advent of information theory [Sha48]. For data rates up to a certain level, called the capacity of the channel, the probability of error can be made arbitrarily small, by combining messages into codewords in a more clever way. On the other hand, this comes at the price of an infinite coding delay, since driving the error probability to zero is the result of using longer and longer codewords.

When channel fading enters the picture, it is not apparent how to generalize the notion of capacity. Messages, or codewords, may be wrongly decoded not only due to present noise, but also due to considerable signal attenuation caused by channel fading. Wireless channels are often classified as either slow-fading or fast-fading, depending on how rapid the channel variations are in comparison to the tolerated delay for communication. In fast fading there is sufficient time to form codewords spanning many fading states, and the “average capacity” over the states, commonly termed ergodic capacity, is the true capacity of the channel. In slow fading, on the other hand, there is no significant channel fading by the transmission time of a codeword. A channel in a fading dip cannot be counteracted by coding, and an outage is declared if the channel strength drops below a given threshold. Performance measures related to the probability of outage are often considered in slow-fading channels.

In delay-sensitive applications, such as voice/video calls or other kinds of real-time services, codewords need to be relatively short. There will always be an associated probability of making decoding errors, both due to channel fading and noise. A useful technique to combine with shorter codewords is that of error control. By adding a few extra check-bits in the end of a message, the receiver can conclude that the decoded message is correct with high probability. If an error has occurred, it can notify the transmitter that the message needs to be retransmitted (assuming there is a reverse link for such control signals).

1.2.1 Wideband Communications

The available frequency band may be substantially larger than the coherence bandwidth of the wireless channel. The combined effect of multipath propagation, with different paths corresponding to different time delays at the receiver, can in this case not be modeled by a common attenuation
and phase-shift over the entire frequency band. Due to the different time delays, consecutively sent messages will be mixed in time in the received signal. The messages can, however, be untwined using equalization—a signal processing technique—at the receiver. The channel variations over the frequency band can even be exploited using additional pre-processing at the transmitter. The joint design of such transmitter and receiver filters (see, e.g., [Sal73]) is often referred to as transceiver design.

Another approach to wideband communications, adopted in several current wireless standards, is orthogonal frequency-division multiplexing (OFDM). It has the desirable effect of converting a single wideband channel into multiple, non-interfering narrowband subchannels. It allows conventional transmission/reception techniques for narrowband communications to be used in the context of wideband communications. While this thesis only explicitly considers a narrowband setting, the use of OFDM makes the results applicable to wideband systems as well. Still, the traditional wideband techniques of transceiver design comprise a central theme of the thesis, although in a different context of narrowband multi-antenna communications.

1.2.2 Multi-Antenna Performance Gains

The use of multiple antennas for transmission and reception can have remarkable positive effects on key performance measures such as channel capacity, outage probability, or the probability of making decoding errors. Imagine first placing additional antennas at the receiving device. The receiver then has access to several signals representing the transmitted signal. If the signals from the receive antennas are combined coherently, compensating for their mismatch in phase, the desired signal terms add up to a greater extent than the undesired noise terms. Such receive processing leads to an overall SNR increase, often referred to as an array gain. The increase in SNR is directly reflected in a decrease in the decoding-error probability in practical communication schemes.

There are two other fundamental performance gains of using multiple antennas for communications over fading channels.

- **Diversity gain:** Multiple antennas can be used as an effective tool to combat the negative aspects of slow channel fading. With sufficiently high SNR on average, decoding errors primarily occur due to the wireless channel being in a fading dip—a valley in Figure 1.3. With multiple antennas on both sides—i.e., a multiple-input multiple-output (MIMO) system—each transmit-receive antenna pair forms a fading channel. One conceptually simple diversity scheme is to always use the best antenna pair for transmission and reception. With $n_T$ transmit antennas and $n_R$ receive antennas, the outage probability drops
CHAPTER 1. INTRODUCTION

by a factor $n_T n_R$ (the number of transmit-receive antenna pairs), assuming independently fading channels. This gain in reliability of communication is the diversity gain.

- **Multiplexing gain:** Wireless MIMO channels in scattering environments are capable of spatial multiplexing. Several data streams can be conveyed in parallel over different propagation paths. Recall the landscape of channel variations in space illustrated in Figure 1.3, caused by multipath propagation. With sufficient scattering in the environment, transmit antennas separated by as little as half a wavelength will give rise to two inherently different landscapes at the receiver. With multiple receive antennas, each transmit antenna will leave its very own spatial signature on the receiving antennas. This means that several information-carrying signals can be sent simultaneously on the transmit antennas and still be separated at the receiver. The performance gain associated with spatial multiplexing, the multiplexing gain, was not recognized until fairly recently [Tel99]; in fast fading, the capacity of the MIMO channel scales linearly with $\min(n_T, n_R)$—the minimum number of antennas on the two sides.

An important remark is that the two multi-antenna performance gains cannot be simultaneously achieved by any communication scheme; the optimal trade-off between the two has been explored in [ZT03]. It can be argued that spatial multiplexing should be favored over diversity schemes in modern wireless systems that already offer diversity in both frequency and time by means of coding [LJ10].

1.2.3 Channel Information

Many communication schemes, especially for MIMO, rely on accurate knowledge of the wireless channel at the receiver. Channel estimation can be performed on the receiver side if, for example, a suitable training signal is transmitted. A requirement for any kind of channel-state information (CSI) at the transmitter, on the other hand, is the presence of a reverse link. In a frequency-division duplexing (FDD) system, there is a reverse link operating in a different frequency band. If the time-variation of the channel is sufficiently slow, it is possible to convey CSI back to the transmitter over the reverse link. In a time-division duplexing (TDD) system, the forward and reverse links use the same frequency band in different time slots. It is then possible to obtain transmitter CSI using channel reciprocity: The wireless channel from a transmit antenna to a receive antenna is the same as that in the reverse direction. By sending training signals on the reverse link, the transmitter can obtain the desired channel estimate for the forward link.
Due to the additional difficulties with providing CSI at the transmitter (CSIT), it is safe to say that CSIT is prone to error. In FDD systems, imperfections arise due to channel estimation, quantization prior to feedback, and an error due to feedback delay. In TDD systems CSIT will become outdated if the channel coherence time is less than the duplexing time. Also, while channel reciprocity holds for the physical channel, it does not hold for the actual circuits used for transmitting/receiving radio waves, and errors follow if such effects are not completely compensated for.

If the channel variation is sufficiently rapid, it is instead reasonable to assume that the transmitter lacks knowledge of small-scale variations of the channel. A long-term, statistical description of the channel, based on large-scale fading parameters, may still be highly useful channel information at the transmitter. This thesis elaborates extensively on how to explore such long-term CSI to improve the performance of MIMO communication schemes.

1.3 Designing MIMO Communication Systems

There is no simple answer to the question of how to utilize a wireless MIMO channel for communications. There are several important facets of selecting a transmit-receive scheme for communications, for example: 1) What kind of channel knowledge can be assumed at the transmitter/receiver; 2) Should focus be set on diversity or multiplexing, or a combination of the two; 3) What are the computational resources available for signal processing at the transmitter/receiver? The latter question touches upon one of the exciting and complicating aspects of MIMO communication schemes: There is in general a trade-off between performance and computational complexity. When transmitting several independent messages simultaneously, for example using spatial multiplexing, optimal receive processing involves taking a joint decision regarding all messages. The number of possible message combinations to consider at the receiver scales prohibitively fast with the number of parallel messages sent. More efficient receive processing can be achieved by taking independent decisions, but leads to a performance loss in general.

Much research effort has been invested in devising computationally efficient communication schemes under various system assumptions, resulting in several quite ingenious solutions. One line of research has focused on space-time block codes (STBC) [TJC99] capable of achieving the full diversity of the MIMO channel without CSIT. Several messages are combined into codewords extending in both time and space (i.e., across the transmit antennas). Special orthogonal STBCs allow efficient receive processing in combination with independent message decoding. However, the reduced computational complexity comes at the price of a poor per-
formance in terms of the possible multiplexing gain. STBCs achieving the diversity-multiplexing trade-off require more complex receive processing [EKP+06, ORBV06].

Several spatial-multiplexing architectures with efficient receive processing were proposed in the late 1990s, and presented under the acronym BLAST (Bell Labs layered space time) [Fos96, WFGV98, FGVW99, FCG+03]. Messages to be transmitted in parallel are assigned to different layers to be independently processed at the receiver, setting the stage for efficient receive processing. The most simple vertical version, V-BLAST, assigns layers to different transmit antennas and manages to achieve the full multiplexing gain of the MIMO channel. The efficiency of the V-BLAST receive processing comes at a price: The possible diversity gain that can be reaped is poor. Optimal joint decoding performs better in this respect but is more computationally demanding.

1.3.1 MIMO Transceiver Design

The basic MIMO transceiver architectures of V-BLAST and STBC were originally proposed assuming perfect CSI at the receiver and no CSIT (i.e., no information regarding small-scale fading). However, new exciting possibilities for MIMO communications arise if it is possible to supply the transmitter with CSI as well. The transmitter can explore channel information to transmit in the directions leading to the receiver via multipath propagation. This is commonly called beamforming.\(^1\) If there is only a large-scale description of the channel available at the transmitter, there might be transmit directions that more commonly than others lead to the receiver. Forming beams in those directions will still be a fruitful strategy. Irrespective of the kind of transmitter CSI, spatial multiplexing can be realized by sending several messages using different beams, instead of simply using the different transmit antennas for the different messages, which is the case for V-BLAST.

The availability of channel information on both sides also provides an opportunity of jointly selecting a transmit and receive strategy based on CSI. This idea comes naturally, since optimal receive processing depends on what is done at the transmitter and vice versa. MIMO transceiver design refers to the joint optimization of the transmit and receive processing parts. While narrowband multi-antenna communications is considered in the thesis, general MIMO systems arise in a variety of different contexts, such as single-antenna wideband systems or wireline DSL systems. The general field of MIMO transceiver design has undergone an exciting journey over the past decade, initiated in the milestone paper [PCL03]. The

\(^1\)In multipath environments, these “beams” are in fact more complex radiation profiles in the angle around the antenna array.
1.3. DESIGNING MIMO COMMUNICATION SYSTEMS

mathematical theory of majorization [MOA11] plays a key role in solving optimization problems related to MIMO transceiver design, and is used extensively in this thesis.

1.3.2 Multi-user Communication Systems

Up to this point, the discussion has revolved around point-to-point communications: a single transmitter and a single receiver. The thesis also considers wireless multi-user systems, in the sense of a single base station (BS) communicating with a number of mobile user devices. This is the setting of a conventional cellular network, where communication in a single cell occurs both in the downlink (from BS to users) as well as in the uplink (from users to BS). Uplink and downlink communication are usually separated using duplexing in time (TDD) or frequency (FDD), as mentioned previously.

There are several different ways to share the wireless channel between the users. In time-division multiple access (TDMA), the users are allocated different, recurring time intervals to transmit. In frequency-division multiple access (FDMA), users transmit simultaneously but on different frequency bands. In the emerging 4G/LTE standard [DPS11], “two-dimensional” time-frequency slots are assigned to different users using the OFDM subchannels, referred to as orthogonal-FDMA (OFDMA). Code-division multiple access (CDMA) uses a somewhat different approach to provide multiple access. In uplink communications, for example, each user is assigned a unique spreading code that is utilized to transmit a narrow-band signal over a wide frequency band, and there are typically many users transmitting at the same time. The BS can distinguish between the users’ signals by correlating with their respective codes. The limiting factor in CDMA systems is interference between users, being perceived as additional noise that reduces the effective SNR. CDMA systems therefore rely on tight power control to reduce overall interference, ensuring that each user transmits with minimum power subject to maintaining its target performance.

The multiple-access techniques of TDMA, FDMA, and OFDMA are designed to let each user communicate with the BS without interference. This allows the previously discussed MIMO communication schemes—assuming a single transmitter and a single receiver—to be used in a multi-user MIMO setting as well. However, it also possible to let several users utilize the same time-frequency slot. Multiple antennas at the BS can then be used for space-division multiple access (SDMA) [RIO96]. In the multi-user MIMO uplink, which is considered in the thesis, the BS can separate the spatial signatures of the users’ transmit antennas. This is the multi-user counterpart of spatial multiplexing considered previously.
1.4 Aim of the Thesis

While the demands on wireless cellular networks are in constant in-
crease, the basic resource to exploit—the frequency spectrum available
for communication—is scarce and hence very expensive to license. This
is spurring research on how to use the available spectrum more and more
efficiently. The use of multiple antennas can play an important role in this
context by allowing for spatial multiplexing and spatial multiple access.
These techniques can be fully utilized when the transmitter and receiver
ends are jointly designed based on available channel information.

The thesis considers MIMO transceiver design for both point-to-point
communication, and multi-user uplink communication with multiple ac-
cess via SDMA. The thesis builds on the recent advances in the field of
MIMO transceiver design [PCL03, JPV07], with majorization theory as a
mathematical tool to address the optimization problems. While perfect
transmitter CSI is often assumed in the context of MIMO transceiver de-
sign, a distinguishing feature of this work is the more practical assumption
of only having a statistical description of small-scale fading available. A
key challenge in this thesis is to pose and efficiently solve MIMO transceiver
design problems in this setting, accounting for the fading nature of wireless
MIMO channels.
Chapter 2

Problem Formulation and Contributions

This chapter gives a more precise description of the MIMO transceiver design problems addressed in the thesis. To this end, the basic input-output signal models are first presented in Section 2.1, along with an account of different statistical models to describe small-scale fading in MIMO channels. Section 2.2 provides a general introduction to MIMO transceivers, as well as transceiver design with different CSI assumptions. In particular, the decision-feedback (DF) transceiver, which is exclusively considered in the remainder of the thesis, is given much focus. The single-user and multi-user transceiver-design problems are presented in Sections 2.2.4 and 2.2.5, respectively. The chapter is concluded with an outline of the contributions of the thesis.

2.1 The Wireless MIMO System Model

Consider a wireless communication system with $n_T$ transmit antennas and $n_R$ receive antennas. The propagation channel from transmit antenna $n$ to receive antenna $m$ on a narrowband carrier is characterized by two effects: the phase-shift and attenuation of the carrier. These are represented by the argument and magnitude, respectively, of the complex baseband-equivalent channel $h_{mn} \in \mathbb{C}$. The scalar channels of the $n_T n_R$ transmit-receive antenna pairs comprise the MIMO channel matrix

$$
H = \begin{bmatrix}
h_{11} & \cdots & h_{1n_T} \\
\vdots & \ddots & \vdots \\
h_{n_R1} & \cdots & h_{n_Rn_T}
\end{bmatrix}. 
$$

(2.1)
In the discrete-time baseband, the transmitted signals \( x(t) \in \mathbb{C}^{n_T} \) and the received signals \( y(t) \in \mathbb{C}^{n_R} \) are related by the narrowband signal model

\[
y(t) = H x(t) + n(t), \tag{2.2}
\]

where \( n(t) \) denotes additive white Gaussian noise (AWGN), modeled as a circular-symmetric complex Gaussian vector \( n(t) \sim \mathcal{C} \mathcal{N}(0, R_n) \) with spatial correlation \( R_n \). The signal model is illustrated in Figure 2.1, showing explicitly that the transmitted signals are mixed at the receiver. The inherent time-varying nature of wireless MIMO channels is not displayed in (2.2), but will be elaborated on in Section 2.1.1.

The signal model (2.2) will be used to represent two different transmission scenarios:

- **Point-to-Point MIMO**: A single transmitter with \( n_T \) antennas, and a single receiver with \( n_R \) antennas.

- **Multi-User MIMO Uplink**: \( K \) simultaneously transmitting users, with user \( k \) having \( n_{Tk} \) antennas (\( n_T = \sum_{k=1}^{K} n_{Tk} \) transmit antennas in total), and a single receiving base station (BS) with \( n_R \) antennas.

While the signal model (2.2) naturally reflects point-to-point MIMO, it can also accommodate the multi-user MIMO uplink model. With user \( k \) transmitting \( x_k(t) \in \mathbb{C}^{n_{Tk}} \) over a MIMO channel \( H_k \in \mathbb{C}^{n_R \times n_{Tk}} \), we may define

\[
\begin{align*}
\{ \mathbf{H} & = [\mathbf{H}_1 \ldots \mathbf{H}_K] \in \mathbb{C}^{n_R \times n_T}, \\
\mathbf{x}(t) & = [\mathbf{x}_1^T(t) \ldots \mathbf{x}_K^T(t)]^T \in \mathbb{C}^{n_T},
\end{align*}
\]

Figure 2.1: Illustration of the linear MIMO signal model.
2.1. **THE WIRELESS MIMO SYSTEM MODEL**

and the MIMO signal model (2.2) transforms into

$$ y(t) = \sum_{k=1}^{K} H_k x_k(t) + n(t). \quad (2.3) $$

This is the multi-user MIMO uplink signal model with $K$ simultaneously transmitting users, which are assumed to be perfectly synchronized in both frequency and time. Reversely, the point-to-point model will also be called single-user MIMO in the following.

The narrowband signal models play an important role in wideband systems as well, as mentioned in Section 1.2.1. In modern wireless standards such as 4G/LTE [DPS11] and WLAN, OFDM is used to effectively establish a number of non-interfering narrowband subcarriers over the available bandwidth [Böl04]. In a multi-user communication system, with different users assigned to the different subcarriers, considerable care needs to be taken to compensate for user frequency offsets and ensuring time-synchronized transmissions, especially in uplink communications (from users to the BS) [MKP07].

2.1.1 **Statistical Modeling of MIMO Channels**

In most practical scenarios a wireless MIMO channel is far from static: The channel is sensitive to antenna movements on both sides, as well as to movements of other objects in the surroundings. The smooth time-evolution of the channel can be modeled and exploited for applications such as channel prediction. For other applications, a detailed model may not be required. A simplistic choice is the quasi-static block-fading model, which regards $H$ as a random process in time that is fixed during a certain coherence time, and then replaced by an independent realization from some distribution.

A canonical statistical model for small-scale fading of MIMO channels is the *i.i.d. Rayleigh model*, in which the entries of the channel matrix $H$ are independent and identically distributed (i.i.d.) circular-symmetric complex Gaussians. Its main motivation is that of a *rich-scattering* propagation environment, with a large number of objects between the transmitter and the receiver that may scatter or reflect the emitted radiation. It is assumed that this gives rise to many physical paths connecting the transmitter and the receiver, being uniformly spread out over a large angle at the transmitter as well as at the receiver. The MIMO channel can then be analyzed in an *angular domain*, where the large number of paths translates into complex Gaussian channel-matrix entries by the central limit theorem, and the uniform path-spread further leads to i.i.d. channel-matrix entries [TV05,Say02]. The argument is based on the assumption of suffi-
cient antennas spacing on both sides: half the wavelength if the angular spread is large.

In a practical scenario, the antennas may not be separated enough for the i.i.d. Rayleigh model to apply. For an elevated base station, the required distance can be up to ten wavelengths due to the poor angular spread [CRL00]. The required separation distance might be much shorter for a hand-held device surrounded by scatterers, but its size may be the limiting factor. This motivates considering MIMO channels with spatial correlation. A common channel model that allows for correlated channel-matrix entries is the separable-correlation Rayleigh model

$$H = R^{1/2}_R Z R^{1/2}_T,$$  \hspace{1cm} (2.4)

which decomposes the channel matrix in terms of a receive correlation matrix $R_R \in \mathbb{C}^{n_R \times n_R}$, a transmit correlation matrix $R_T \in \mathbb{C}^{n_T \times n_T}$, and $Z \in \mathbb{C}^{n_R \times n_T}$ having i.i.d. $\mathcal{CN}(0,1)$ entries. As illustrated in Figure 2.2, the model asserts that channel correlations arise from local scattering at the antenna arrays. For example, $R_R$ is the common correlation matrix (up to a constant) for all columns of $H$. And since the same correlations appear irrespective of which transmit antenna is used, these correlations are attributed to the receive side. The separable-correlation model has been verified in several measurement campaigns, for example [YBO+01, KSP+02, CLW+03].

Even if there are local scatterers and sufficient antenna spacing at both sides, the i.i.d. Rayleigh model might still not apply: Keyhole effects may arise if the transmitter and receiver are far apart and separated by a non-scattering environment, causing an effective rank reduction of the channel matrix. This shortcoming of the Rayleigh models is addressed by the double-scattering model

$$H = R^{1/2}_R Z_R R^{1/2}_S Z_S R^{1/2}_T,$$  \hspace{1cm} (2.5)

with receive and transmit correlation matrices as before, $Z_R \in \mathbb{C}^{n_R \times n_S}$ and $Z_S \in \mathbb{C}^{n_S \times n_T}$ having i.i.d. $\mathcal{CN}(0,1)$ entries, and an intermediate scattering
correlation matrix $R_S \in \mathbb{C}^{n_S \times n_S}$. The model was proposed in [GBGP02] and verified by ray-tracing techniques, while measurements confirming the existence of keyhole effects can be found in [CFGV02] and [ATM03].

The two models (2.4) and (2.5) will be used exclusively in the thesis to model fading point-to-point MIMO channels.\(^1\) For completeness, however, we shall also mention the widely adopted Rician model. When there is a direct line-of-sight between the transmitter and receiver, or a small number of prominent paths, the channel matrix will not be zero-mean. If $\text{vec}(H)$ denotes the $(n_R n_T \times 1)$-vector obtained by stacking the columns of $H$, the Rician model reads

$$\text{vec}(H) \sim \mathcal{CN}\left(\text{vec}(\hat{H}), R_H\right), \tag{2.6}$$

where $\hat{H} = \mathbb{E}[H]$ is the channel mean, and $R_H$ denotes the covariance matrix. While $R_H$ can account for arbitrary channel correlations, it may also be impractical to use due to its large size ($n_R^2 n_T^2$ entries).

### 2.1.2 Noise Pre-Whitening

The point-to-point MIMO and the multi-user MIMO uplink signal models in (2.2) and (2.3), respectively, both assume a single receiver. It is then not necessary to explicitly consider the case of a spatially correlated noise term. Potential noise correlation can be dealt with using a noise pre-whitening procedure, which includes the following two steps:

1. The noise correlation $R_n$ is perfectly estimated at the receiver;

2. The received vector $y(t)$ is processed as

$$y_w(t) = R_n^{-1/2} y(t) = H_w x(t) + n_w(t),$$

producing the desired uncorrelated noise vector $n_w(t) = R_n^{-1/2} n(t) \sim \mathcal{CN}(0, I)$, and a modified channel matrix $H_w = R_n^{-1/2} H$.

In the following, we shall assume that these steps have already been performed, and simply use the signal models under the assumption $n(t) \sim \mathcal{CN}(0, I)$. Note that the noise pre-whitening step effectively contributes to the receive-correlation matrix in the statistical channel models (2.4) and (2.5).

\(^1\)Two related and important contributions to MIMO channel modeling that will not be considered are the virtual channel representation [Say02] and the Weichselberger model [WHOB06].
2.2 MIMO Transceiver Design

The spatial-multiplexing capabilities of wireless MIMO channels allow several data streams to be simultaneously conveyed over a single point-to-point link. The transmit-receive processing steps of first converting data streams into $n_T$ signals at the transmitter, and then recovering the data streams from $n_R$ signals at the receiver, is a description of a MIMO transceiver. The art of optimizing the performance of MIMO transceivers, based on the kind of channel information available at the two ends, is commonly termed MIMO transceiver design and is a central theme in the thesis. This section introduces several MIMO transceivers, with special focus on the decision-feedback transceiver mainly considered in the remainder of the thesis. We will initially consider the point-to-point setting, and dedicate Section 2.2.5 to transceiver design in the multi-user MIMO uplink.

A central notion for the following discussion is that of a data stream, carrying information to be conveyed over the communication channel. A data stream is represented as a complex-valued random signal $s(t)$. In case of an uncoded data stream, $s(t)$ may only take on values that are symbols in a finite constellation $\mathcal{X} \subset \mathbb{C}$. We also assume that all possible symbols in $\mathcal{X}$ are equally likely, and that statistically independent symbols are transmitted over time. The size $|\mathcal{X}|$ of the constellation then determines the number of information bits $\log_2(|\mathcal{X}|)$ that are represented by a symbol, as illustrated in Figure 2.3. The concept of a coded data stream is a bit more involved. For simplicity, we may assume that $s(t)$ is comprised of a sequence of independent codewords, each being a block of consecutive symbols. Similarly as for uncoded data streams, the number of information bits represented by a codeword is determined by the total number of possible codewords, assuming that these are equally likely.
2.2. MIMO TRANSCEIVER DESIGN

2.2.1 Channel Information

The very idea of transceiver design is to adapt a MIMO transceiver to the channel. A natural question is then what kind of information about the channel that is available at the transmitter/receiver. For modeling purposes, such channel-state information (CSI) can be categorized as follows.

- **Perfect CSI**: At any time instant, the channel matrix $H$ is completely known.

- **Imperfect CSI**: The channel matrix is expressed as $H = \hat{H} + \Delta$ in terms of a known channel estimate $\hat{H}$ and an unknown error $\Delta$. In a deterministic model, $\Delta$ belongs to a known uncertainty region $\mathcal{R} \subset \mathbb{C}^{nR \times nT}$. In a statistical model, $\Delta$ is random and its distribution is known. Typically, $H$ is then modeled using a Rician distribution in (2.6) with mean $\hat{H}$.

- **Long-Term Statistical CSI**: Only a statistical model of the small-scale fading is known, for example one of the Rayleigh models, the Rician model, or the double-scattering model (all discussed in Section 2.1.1).

It is commonly assumed that the receiver in a point-to-point MIMO system has access to perfect CSI. Ensuring that the channel can be accurately estimated at the receiver is in general a high priority. Firstly, it enables coherent detection, which is necessary for the use of good constellations. Secondly, accurate CSI is also a precondition for spatial separation of data streams, used in spatial multiplexing and uplink SDMA. Perfect CSI at the receiver is assumed throughout the thesis.

Following the discussion in Section 1.2.3, providing accurate transmitter CSI is more problematic in fading channels. In an FDD system, the combined effects of channel estimation, quantization and error due to feedback delay can be accounted for using an imperfect-CSI model. For a rapidly varying channel, it is reasonable to assume that the transmitter is only aware of long-term, statistical CSI. Such information may be obtained by direct estimation on the reverse link instead of via feedback, although a transformation compensating for the frequency offset is required [BO01]. In a TDD system where channel reciprocity can be exploited, an imperfect-CSI or even perfect-CSI model may apply if the channel variations are sufficiently slow. However, if the channel coherence time is less than the duplexing time, any estimated short-term CSI will become outdated. Long-term, statistical CSI may then be estimated instead.

2.2.2 Single-Stream Beamforming

As an introduction to MIMO transceiver design, we shall illuminate a few key concepts in a simple setting. Consider transmitting a single data stream...
using multiple antennas by means of a transmit beamvector \( \mathbf{p} \in \mathbb{C}^{n_T} \). The transmitted signal is simply given by

\[
\mathbf{x}(t) = \mathbf{p}s(t),
\]

and the received signal is

\[
\mathbf{y}(t) = \mathbf{H}\mathbf{p}s(t) + \mathbf{n}(t).
\]

The receiver can similarly use a receive beamvector \( \mathbf{w} \) to form a signal estimate \( \hat{s}(t) = \mathbf{w}^H\mathbf{y}(t) \), or explicitly,

\[
\hat{s}(t) = \mathbf{w}^H\mathbf{H}\mathbf{p}s(t) + \mathbf{w}^H\mathbf{n}(t).
\] (2.7)

In case \( s(t) \) is an uncoded data stream, the receive processing is completed by a separate detection step; the output of the receiver \( \hat{s}(t) \in \mathcal{X} \) is obtained by simply mapping \( \hat{s}(t) \) onto the nearest constellation point. In case of a coded data stream, \( \hat{s}(t) \) is similarly obtained by feeding \( \hat{s}(t) \) through a decoder, resulting in a delay by the length of a codeword.

A fundamental performance measure for the single-input single-output (SISO) AWGN signal model in (2.7) is the signal-to-noise ratio (SNR)

\[
\gamma = \frac{\mathbb{E}[|\mathbf{w}^H\mathbf{H}\mathbf{p}s(t)|^2]}{\mathbb{E}[|\mathbf{w}^H\mathbf{n}(t)|^2]} = \frac{|\mathbf{w}^H\mathbf{H}\mathbf{p}|^2}{\mathbf{w}^H\mathbf{w}},
\]

assuming spatially uncorrelated noise \( \mathbf{n}(t) \sim \mathcal{CN}(0, \mathbf{I}) \) and a data stream normalized as \( \mathbb{E}[|s(t)|^2] = 1 \). In this context, transceiver design refers to the optimization of the beamvectors \( \mathbf{p} \) and \( \mathbf{w} \). An SNR-maximizing receive beamvector can be found using the Cauchy-Schwarz inequality

\[
|\mathbf{w}^H\mathbf{H}\mathbf{p}|^2 \leq \mathbf{w}^H\mathbf{w} \cdot \mathbf{p}^H\mathbf{H}^H\mathbf{H}\mathbf{p},
\]

with equality if and only if \( \mathbf{w} \propto \mathbf{H}\mathbf{p} \). Hence, the matched filter

\[
\mathbf{w}_{MF} = \frac{\mathbf{H}\mathbf{p}}{\mathbf{p}^H\mathbf{H}^H\mathbf{H}\mathbf{p}}
\] (2.8)

renders an effective signal model

\[
\hat{s}(t) = s(t) + \hat{n}(t),
\]

where \( \hat{n}(t) = \mathbf{w}_{MF}^H\mathbf{n}(t) \), with corresponding SNR

\[
\gamma_{MF} = \mathbf{p}^H\mathbf{H}^H\mathbf{H}\mathbf{p}.
\] (2.9)

It is evident from (2.8) that optimal receive beamforming requires knowledge of the effective channel \( \mathbf{H}\mathbf{p} \), or perfect CSI in combination with the transmit beamvector.
With the optimal receive beamvector $\mathbf{w}_{MF}$ given by (2.8), the solution to the joint transmitter-receiver optimization problem is given by maximizing $\gamma_{MF}$ with respect to the transmit beamvector $\mathbf{p}$. We assume that the total transmit power $P_{\text{max}}$ over the antennas is fixed:

$$P_{\text{max}} = \mathbb{E} \left[ ||\mathbf{x}(t)||^2 \right] = \mathbb{E} \left[ ||\mathbf{p}s(t)||^2 \right] = \mathbf{p}^H \mathbf{p}.$$  

In what follows, we shall contrast different assumptions on CSI at the transmitter, as well as different statistical channel models.

- **Perfect CSI:** The optimal $\mathbf{p}$ is readily found using the inequality

$$\gamma_{MF} = \mathbf{p}^H \mathbf{H}^H \mathbf{H} \mathbf{p} \leq \lambda_1 \cdot \mathbf{p}^H \mathbf{p} = \lambda_1 P_{\text{max}},$$

where $\lambda_1$ is the largest eigenvalue of $\mathbf{H}^H \mathbf{H}$. Equality is obtained by aligning $\mathbf{p}$ with the corresponding eigenvector. The maximal SNR that can be achieved is proportional to the allowed transmit power.

- **Statistical CSI – i.i.d. Rayleigh:** Assuming instead that only long-term statistical CSI is available at the transmitter alters the nature of the transmitter design problem. Considering the i.i.d. Rayleigh model, with $\mathbf{H}$ having i.i.d. $\mathcal{CN}(0,1)$ entries, the SNR $\gamma_{MF}$ in (2.9) is a random variable with distribution

$$\gamma_{MF} \sim \mathbf{z}^H \mathbf{z} \cdot P_{\text{max}}, \quad (2.10)$$

where $\mathbf{z} \in \mathbb{C}^{nR}$ is a random vector with i.i.d. $\mathcal{CN}(0,1)$ elements, and $P_{\text{max}} = \mathbf{p}^H \mathbf{p}$ as before. In this case, it is impossible to alter the SNR distribution by means of a transmit beamvector $\mathbf{p}$ if the transmit power is kept fixed. Hence, there is no transmitter design problem, and we may simply use a single, arbitrarily selected antenna for transmission!

- **Statistical CSI – Correlated Rayleigh:** Interestingly, a different conclusion is reached under the correlated Rayleigh model $\mathbf{H} = \mathbf{R}_R^{1/2} \mathbf{Z} \mathbf{R}_T^{1/2}$ in (2.4). The SNR distribution in (2.10) generalizes to

$$\gamma_{MF} \sim \mathbf{z}^H \mathbf{R}_R \mathbf{z} \cdot \mathbf{p}^H \mathbf{R}_T \mathbf{p}. \quad \text{While the random component} \quad \mathbf{z}^H \mathbf{R}_R \mathbf{z} \quad \text{is independent of} \quad \mathbf{p}, \quad \text{there is now a power gain available by aligning} \quad \mathbf{p} \quad \text{with the transmit correlation matrix} \quad \mathbf{R}_T. \quad \text{Under a transmit power constraint, this procedure is the same as for perfect CSI by replacing} \quad \mathbf{H}^H \mathbf{H} \quad \text{by} \quad \mathbf{R}_T.$$

2.2.3 MIMO Transceivers for Spatial Multiplexing

A transmit strategy related to single-stream beamforming is to simultaneously transmit multiple data streams $s_1(t), \ldots, s_{n_L}(t)$ with associated transmit beamvectors $p_1, \ldots, p_{n_L}$. The transmitted signal can then be written as

$$x(t) = p_1 s_1(t) + \ldots + p_{n_L} s_{n_L}(t).$$

The matrix $P = [p_1 \ldots p_{n_L}]$ formed by beamvectors is called a linear precoder, and the transmitted signal $x(t)$ can be compactly expressed as

$$x(t) = Ps(t),$$

where $s(t) = [s_1(t) \ldots s_{n_L}(t)]^T$ is the vector of data streams. The data streams are assumed to be independently encoded, and normalized as $\mathbb{E} [s(t)s(t)^H] = I$.

Assuming that the precoder and the MIMO channel are both known at the receiver, the optimal detector in terms of minimizing the probability of making an error is given by the maximum-likelihood criterion [DEC03, KP98]

$$\hat{s}(t) = \arg\min_s ||y(t) - HPs||^2. \quad (2.11)$$

The search in $s$ is carried out over the set of possible transmit vectors. With a common constellation $\mathcal{X}$ for all data streams, the number of symbol vectors are $|\mathcal{X}|^{n_L}$. The exponential scaling in the number of data streams may render the detection problem computationally demanding, and large constellations induce the same effect. The sphere-decoder implementation can be used to reduce number of symbol-vectors to search over [DEC03], but the average complexity of the detector in i.i.d. Rayleigh fading is still exponential in $n_L$ [JO05].

Efficient Receivers for Spatial Multiplexing

In the late 1990s, transceiver architectures for wireless MIMO channels were subject to intense research. This research was spurred by the insight of the spatial-multiplexing capabilities of wireless channels [Tel99, FG98]. The V-BLAST architecture [WFGV98, FGVW99] comprised an important contribution, showing that the benefits of spatial multiplexing could be reaped with efficient signal processing at the receiver.

The decrease in receiver complexity stems from independent detection of the simultaneously transmitted symbols. The process of forming a predetection signal estimate $\bar{s}(t)$ of $s(t)$ is often referred to as equalization. The receive beamvector $w$ for single-stream beamforming has its natural counterpart in a linear equalizer $W \in \mathbb{C}^{RR \times n_L}$ in the sense that a
signal estimate is formed as \( \bar{s}(t) = W^H y(t) \). One possible generalization of the matched filter \( w_{\text{MF}} \) is the zero-forcing (ZF) linear equalizer \( W_{\text{LIN}} = HP\left(P^H H^H HP\right)^{-1} \), which creates an effective signal model

\[
\bar{s}(t) = s(t) + \hat{n}(t), \quad (2.12)
\]

where \( \hat{n}(t) = W_{\text{LIN}}^H n(t) \). This channel inversion is only possible if the receiver has perfect knowledge of the effective channel \( HP \), and if the number of data streams \( n_L \) does not exceed the rank of the channel. The particular choice \( W_{\text{LIN}} \) can be shown to maximize the SNR on each link without trade-off, subject to achieving the zero-forcing signal model in (2.12). With simple signal processing, the MIMO channel is converted into parallel SISO channels, and independent detection can be performed on each subchannel.

The linear receiver and the ML receiver represent two extremes in the trade-off between computational complexity and performance; linear equalization and independent symbol detection is readily implemented, but often at the expense of a significantly increased probability of making detection errors. The receive processing proposed in V-BLAST combines linear equalization with a technique called successive interference cancellation (SIC). While novel in this context, this combination of equalization techniques was already known in the field of MIMO transceiver design as decision feedback (DF) equalization [GC01]. We describe the ZF-DF receive processing next in terms of the QL decomposition \( QL = HP \), where \( Q \in \mathbb{C}^{n_R \times n_L} \) has orthonormal columns and \( L \in \mathbb{C}^{n_L \times n_L} \) is lower triangular with positive diagonal entries.

1. **Linear equalization**: The first step, also called interference nulling, is performed using the linear equalizer \( W_{\text{DF}} = QD_L^{-1} \), where \( D_L \) is a diagonal matrix with the same diagonal as \( L \). The processed signal becomes

\[
W_{\text{DF}}^H y(t) = D_L^{-1} L s(t) + D_L^{-1} Q^H n(t).
\]

In this manner, interference nulling is performed in close connection with the natural detection order from \( s_1(t) \) to \( s_{n_L}(t) \); component \( k \) of the signal term \( D_L^{-1} L s(t) \),

\[
[D_L^{-1} L s(t)]_k = \frac{l_{k,1}}{l_{k,k}} s_1(t) + \ldots + \frac{l_{k,k-1}}{l_{k,k}} s_{k-1}(t) + s_k(t), \quad (2.13)
\]

contains interference only from symbols \( s_1(t) \) to \( s_{k-1}(t) \), which will all be detected prior to symbol \( k \).
2. *Successive Interference Cancellation*: By detecting the symbols sequentially, remaining interference in (2.13) can be determined and subtracted, provided that the detection of preceding symbols was successful. After interference nulling and cancelling, a ZF signal model \( \bar{s}(t) = s(t) + \hat{n}(t) \) is obtained, where \( \hat{n}(t) = D_L^{-1}Q^Hn(t) \).

Similar to the linear ZF receiver, the ZF-DF receiver establishes \( n_L \) SISO AWGN subchannels. As opposed to the linear receiver, these are truly parallel in the sense that the noise terms are statistically independent (\( \hat{n}(t) = D_L^{-1}Q^Hn(t) \) has diagonal covariance). The ZF-DF subchannels can also be used for coded data streams. In this case, entire codewords are decoded and used in the interference-cancellation step [FCG03].

The DF equalizer does not provide any SNR benefit over the linear equalizer for the first subchannel to be processed. However, with successful interference cancellation there is an increasing benefit with the placement in the detection order; the last subchannel sees no interference, and its SNR coincides with that of the matched filter in (2.9) for single-stream transmission.

An inherent drawback of SIC is *error propagation*: If a detection error is made, interference cancellation will fail for the remaining subchannels and the probability of making additional errors will significantly increase. There are several techniques available to mitigate the effect of error propagation. The original description of the V-BLAST receiver included a greedy rule to obtain a detection order to mitigate error propagation [WFGV98]: In each stage of the equalization process, the data stream with the lowest probability of error is selected. Greedy algorithms do not in general fulfill a global optimality criterion, but it was verified that the algorithm maximizes the weakest subchannel SNR over all detection orders. The original implementation had \( \mathcal{O}(n^4_L) \) complexity, as compared to \( \mathcal{O}(n^3_L) \) for unordered V-BLAST (DF equalization), but algorithmic advances on the V-BLAST ordering have resulted in \( \mathcal{O}(n^2_L) \) implementations [Has00,SX09,ZCLG11]. Another heuristic greedy algorithm was presented in [WBR01], equalizing the subchannel SNRs from the other end—by minimizing the maximum SNR—leading to reduced complexity. The problem of finding an optimal order based on a general performance objective is in general hard, since the number of detection orders scales as \( n_L! \) with the number of data streams \( n_L \). Two examples in the literature considering an exhaustive search to solve the ordering problem are [PV04] and [ALA10]. The former optimizes the joint symbol error probability (JEP), which is the probability of making at least one detection error, while the latter optimizes a weighted sum of symbol error probabilities.

There are also other means to reduce error propagation. It is for example possible to use iterative detection at the receiver. Soft symbol estimates, weighted based on their reliability, can be used in the interference-
2.2. MIMO TRANSCEIVER DESIGN

Cancellation step. By updating the soft estimates in an iterative manner, enhanced performance can be achieved [Hon09]. Another remedy to the problem of error propagation is to use coded data streams [FCG+03]; using slightly stronger codes for data streams decoded early on reduces error propagation. If there is an error-control mechanism available, it is even possible to make sure not to use feedback when decoding fails [PWJH11].

Remark 2.1. The ZF equalizers for both linear and DF receivers described above are the optimal solutions to the receiver design problem of optimizing subchannel SNR subject to a ZF signal model on the form in (2.12) [PJ06, SD08b]. For both receivers, the linear-equalization step for each data stream amounts to projecting the received signal \( y(t) \) onto a subspace perpendicular to that spanned by its interferers. Unfortunately, the signal component in the interference subspace is lost as well. Part of this may be recovered by relaxing the constraint on zero interference, and minimizing the MSE \( \mathbb{E}[(\hat{s}_k(t) - s_k(t))^2] \) instead. Minimum-MSE (MMSE) equalizers provide a more delicate trade-off between noise and interference suppression, which leads to enhanced performance (see, e.g., [JVL11]).

2.2.4 Transceiver Design for Spatial Multiplexing

If there is channel information available on the transmitter side, linear precoding is an efficient and effective means to enhance the quality of communication over a MIMO link. The effective MIMO channel is altered to make better use of the spatial dimensions. To enjoy linear precoding to its full extent, it is necessary to design the transmitter and receiver jointly, based on available channel information.

The previous section presented the optimal linear and DF receivers under a ZF constraint. Since these are functions of the linear precoder used, joint transmitter-receiver design is performed in practice by focusing solely on optimizing the precoder. It is common to maximize system performance subject to a constraint on the total transmit power over the antennas:

\[
\mathbb{E} \left[ ||x(t)||^2 \right] = \text{tr} (P^H P) \leq P_{\text{max}}.
\]

With a general cost function \( \mathcal{F}(\gamma) \) representing system performance with \( n_L \) parallel AWGN subchannels having SNRs \( \gamma = (\gamma_1, \ldots, \gamma_{n_L}) \), the optimization problem for the ZF-DF receiver reads

\[
\begin{align*}
\text{minimize} & \quad \mathcal{F}(\gamma) \\
\text{subject to} & \quad \mathbf{Q} \mathbf{L} = \mathbf{H} \mathbf{P} \\
& \quad \gamma = (l_1^2, \ldots, l_{n_L}^2) \\
& \quad \text{tr} (P^H P) \leq P_{\text{max}}
\end{align*}
\]  

(2.14)
CHAPTER 2. PROBLEM FORMULATION

in terms of the QL decomposition appearing in the description of the ZF-DF receiver. Similarly as for transmit beamforming, different assumptions on transmitter CSI lead to quite different optimization problems. While the formulation above naturally applies to perfect CSI, it may also represent the transmitter design problem assuming long-term statistical CSI. In this case, the quantities $Q, L, H,$ and $\gamma$ are all random and $F(\gamma)$ represents average performance over the channel fading statistics via the probability distribution of $\gamma$. The same approach can be taken for a statistical model of short-term, imperfect CSI. A deterministic model, on the other hand, is commonly addressed using a worst-case approach: The objective to minimize is the maximum cost over all possible estimation errors $\Delta$ in the uncertainty region $\mathcal{R}$.

The optimization problem (2.14) can be supplemented with additional constraints. It might for example be desirable to introduce per-antenna power constraints, or add quality of service (QoS) constraints on individual subchannels. Alternatively, the entire optimization problem can be reformulated as the minimization of the total transmit power, subject to QoS constraints. While this thesis only treats optimization problems on the form in (2.14), it should be noted that related formulations can be addressed with similar techniques. These alternative forms of the precoder-optimization problem are covered in detail in [PJ06] for the case of perfect CSI at the transmitter.

An important contribution to MIMO transceiver design was presented in [PCL03], optimizing linear transceivers under the assumption of perfect CSI at the transmitter. The mathematical theory of majorization [MOA11] was explored to reformulate the transceiver design problem as a convex optimization problem. The plethora of different performance measures appearing in the literature could be addressed in a unified way. This approach was later extended by [JPV07] and [SD08a] to include DF transceivers with perfect transmitter CSI. The optimal precoders were expressed using a new matrix decomposition: the generalized triangular decomposition (GTD) [JHL08]. With a CSI-aware transmitter and optimal linear precoding, DF equalization is often superior in performance compared to the linear equalizer for a wide range of performance measures [PJ06]. However, it was recently realized that this gain vanishes when a total data rate is optimally allocated to the data streams [BPO09, WCV10]. In this case, the optimal DF transceiver does not use interference cancellation, and is hence reduced to a linear transceiver.

The common assumption on perfect CSI at the transmitter is an idealization, and several lines of research have addressed robustness, either by optimizing worst-case performance [WP09, PPPL06], or using a statistical model of imperfect short-term CSI [ZPO08]. MIMO transceiver design assuming long-term, statistical CSI at the transmitter was addressed in [KS04c] for the linear equalizer, and in [PV04, KL08, LZW09] for the DF
2.2. MIMO TRANSCEIVER DESIGN

While [PV04] and [KL08] minimize different error probabilities for spatially uncorrelated channels, the design in [LZW09] targets expected mean square error (MSE) and a channel model allowing spatial correlation on the transmitter side only. Another approach that explicitly deals with limited feedback using a Grassmannian codebook was developed by [LH05] for linear receivers, and extended to the DF equalizer in [SD08b]. Limited feedback DF was also considered in [JV08a]. The proposed MIMO transceiver included an optimal-ordering step at the receiver, and the optimal detection order was fed back to the transmitter.

This thesis focuses on MIMO transceiver design for the ZF-DF receiver in spatially correlated MIMO channels, assuming long-term statistical CSI at the transmitter. A prerequisite for transceiver design is the availability of expressions representing system performance. For a fixed detection order and disregarding error propagation, common available expressions for scalar AWGN channels can be used to compute instantaneous performance. However, the average performance over time depends on the probability distribution of the channel via the statistical channel model adopted. The SNR distributions for the linear receiver in transmit-correlated Rayleigh fading was presented in [GHP02]. Accounting for receive correlation is more involved, but [KS04c] provided an SNR characterization in terms of the moment-generating function, allowing error probabilities to be computed by means of numerical integration. Bounds on the achievable sum rate in the same scenario were recently proposed in [MZR11]. For the ZF-DF equalizer, an SNR characterization for the transmit-correlated Rayleigh case was presented in [LZW09], and the average MSE was considered for transceiver design. This thesis contributes by presenting expressions for a wide range of performance measures for the ZF-DF receiver in the presence of receive correlation as well, and the correlated double-scattering model is also considered.

2.2.5 Transceiver Design in the Multi-User Uplink

The ZF-DF receiver applies to spatial multiplexing and spatial multiple-access in the multi-user MIMO uplink as well; from the receiver’s point of view, it is of minor importance from where the transmitted signals originate, apart from synchronization and coding. The multi-user uplink is illustrated in Figure 2.4. In this setting, user \( k \) employs \( n_{Lk} \) independent data streams \( s_{k}(t) = (s_{k,1}(t), \ldots, s_{k,n_{Lk}}(t)) \). The data streams are processed by a linear precoder \( P_k \in \mathbb{C}^{n_{Lk} \times n_{Lk}} \) to form the transmitted signal vector

\[
x_k(t) = P_k s_k(t).
\]

In terms of a concatenated signal vector \( s(t) = [s_1(t)^T \ldots s_K(t)^T]^T \), and a block-diagonal precoder \( P = \text{diag}(P_1, \ldots, P_K) \), the effective signal model
can be written as
\[
y(t) = \sum_{k=1}^{K} H_k P_k s_k(t) + n(t) \\
= HPs(t) + n(t),
\]
where \( H = [H_1 \ldots H_K] \). Since the multi-user MIMO uplink model can be written in single-user form, the description of the ZF-DF equalizer directly applies in the multi-user setting. Hence, the ZF-DF receiver decomposes the multi-user MIMO uplink into non-interfering subchannels as well. This point is illustrated to the left in Figure 2.5 in a two-user uplink with two subchannels per user, in which the detection order is implicit; an assignment of subchannels to placements in the detection order is illustrated to the right in Figure 2.5, and affects the SNRs on the subchannels.

In the multi-user MIMO uplink, [YRBC04] addressed the problem of jointly optimizing linear precoders for all users to achieve the sum capacity, assuming fixed channels, perfect CSI available everywhere and individual user transmit-power budgets. It was shown that the optimal set of precoders could be found iteratively, using the waterfilling solution of the corresponding single-user setup [Tel99]. More realistic assumptions on CSI for wireless channels was considered in [SU07]. Assuming only statistical CSI at the transmitters, means to obtain sum-capacity achieving precoders were presented. The corresponding capacity region was discussed in [JVG01].

The sum-capacity achieving precoders from [YRBC04] and [SU07] may lead to different rates for the users depending on the receiver. There is in fact a sum-capacity surface, and the extremal points on this surface can be attained using different orders of a DF equalizer [VG97]. Means to find a fair operating point on this surface was presented by [MMK09], and it can realized using time-sharing or rate-splitting techniques. MIMO transceiver design in the uplink has also been addressed using linear equal-
2.3 Outline of Thesis Contributions

2.3.1 Single-User Transceiver Design

The first part of the thesis (chapters 3 to 5) considers MIMO transceiver design for the ZF-DF equalizer based on long-term statistical CSI at the transmitter. The main contributions to single-user transceiver design are summarized as follows.

- We present a design framework for precoder optimization covering a wide range of relevant performance measures. It applies to the correlated Rayleigh-fading and double scattering channel models. In particular, we propose a novel convex formulation for optimizing a general objective that can be efficiently implemented using standard algorithms of convex optimization; the number of constraints and variables is quadratic in the number of data streams $n_L$.

- Transmitter design with equal-rate subchannels is given special attention. Common performance measures for such systems in transmit-correlated Rayleigh fading are analytically analyzed, resulting in a
reduced-complexity formulation; the number of constraints and variables is linear in the number of data streams $n_L$.

- A general class of MSE-based performance measures is studied in detail, and we show that there is no need to resort to standard algorithms of convex optimization. An exact solution to the power-allocation problem leading to the optimal precoder can be solved using a simple algorithm that identifies the convex hull of a set of points in $\mathbb{R}^2$.

- We derive explicit approximate expressions for a variety of performance measures under the separable-correlation Rayleigh and double-scattering channel models.

Outline of the Presentation

The material is organized as follows. Chapter 3 presents the optimization framework, particularized to transmit-correlated Rayleigh fading. We evaluate the performance of optimized designs in a 4-by-4 MIMO system. We investigate numerically the impact of channel correlation at the transmitter, subchannel rate allocation, optimal detection ordering, and the suboptimality of MSE-based designs.

Chapter 4 extends the applicability of the framework by considering more generally correlated Rayleigh and double scattering MIMO channel models. For these channel models, explicit expressions for ZF-DF subchannel performance measures are not readily available in the literature, as argued in Section 2.2.4. A somewhat involved characterization of the subchannel SNRs in terms of the moment-generating function was presented in [KS04c]. While these results can be extended to the ZF-DF transceiver, we adopt an alternative approach in this chapter. We present accurate approximate SNR distributions for both correlated Rayleigh and double-scattering models that admit simple expressions for a variety of performance measures.

In Chapter 5, we include a more general investigation on the underlying structure of optimization problems that arise in MIMO transceiver design under different assumptions on transmitter CSI. In particular, we establish two new connections between MIMO transceiver design and other disciplines: two-dimensional convex-hull algorithms [MA79], and optimization with submodular constraints [Fuj05]. We exemplify that precoding problems with strict per-antenna power constraints can be solved directly by applying algorithms from these disciplines.

Chapters 3 and 5 are based on the following articles:
2.3. OUTLINE OF THESIS CONTRIBUTIONS


Remark 2.2. Section 5.1 is a shared contribution with Dr. Svante Bergman. In the process of writing this thesis, the material has been revised in part, and extended in its applicability in Section 5.1.2. Chapter 4 and Section 5.2 are preliminary results of current research that have not yet been submitted for publication.

A related contribution to MIMO transceiver design that has not been included in the thesis is the following:


2.3.2 Multi-User Transceiver Design

The second part of the thesis is confined to Chapter 6. We consider transceiver design in a multi-user MIMO uplink with a ZF-DF receiver, jointly optimizing precoders and a decoding order based on long-term statistical CSI. The main contributions can be summarized as follows.

- We characterize the joint distribution of the subchannel SNRs in the MIMO uplink under ZF-DF equalization where each user’s channel either follows a correlated Rayleigh or a double-scattering channel model. We conclude that each user’s subchannel SNRs are statistically invariant to the other users’ choice of precoder; joint precoder optimization in the uplink reduces to solving single-user precoding problems. As a consequence, general user-performance regions have the shape of a polyblock—a finite union of hyper-rectangles, one for each possible decoding order—or a convex polytope if time-sharing is allowed.

- We investigate the decoding ordering problem under a transmit-correlated Rayleigh-fading channel model. Realizing that such or-
dering problems can be cast into the framework of linear assignment problems, efficient algorithms can be employed to solve the ordering problem for a variety of system performance measures. Using majorization theory [MOA11], we also identify relevant cases in which the optimal order can be directly obtained.

- We propose the use of alternating optimization to address joint precoder–decoding-order optimization in a computationally efficient manner. This suboptimal, iterative approach is observed numerically to converge fast and perform close to optimally.

The approach to decoding ordering in the thesis contrasts conventional orderings that dynamically adapt to the fading channels states, for example [WBR+01, PV04, ALA10, WFGV98, JV08a]; the decoding order remains fixed until long-term channel statistics change. While an ordering policy that optimally exploits short-term CSI should perform better than a fixed ordering, there are several reasons to consider the latter. Firstly, common CSI-based orderings only exploit current CSI (and not past CSI as well), and it will be demonstrated in Chapter 6 that such orderings can be inferior to long-term ordering from a user-fairness perspective over a longer time scale. Secondly, with sufficient user spread in terms of path-loss, a short-term ordering may be virtually static, coinciding with an optimal long-term order. Thirdly, long-term ordering enables DF equalization to be combined with channel coding with codewords extending over several channel-coherence intervals, which is appropriate for fast-fading channels. Lastly, since short-term ordering receivers are hard to model, a well-performing set of precoders may be obtained by joint optimization with a fixed decoding order, although a dynamic order may be used instead in the end.

The contributions to MIMO transceiver design in the multi-user MIMO uplink are based on the following articles:


Chapter 3

A Framework for Precoder Optimization

This chapter lays the foundation for precoder optimization for the ZF-DF transceiver based on long-term, statistical CSI at the transmitter. The main focus is to obtain simple convex formulations that can be solved using efficient standard algorithms of convex optimization. The framework for optimization is formulated under a Rayleigh-fading channel model with spatial correlation on the transmitter side only. This setting is suitable for illustrating the key points of the framework. The next chapter generalizes the results to allow for more generally correlated Rayleigh-fading and double-scattering channel models.

3.1 System Model

The DF transceiver for spatial multiplexing was introduced in Section 2.2.3 of the preceding chapter, and is illustrated here in Figure 3.1.\(^1\) The context

\[^1\text{Note that in this chapter, as well as in the remainder of the thesis, the time index of time-dependent signals is dropped.}\]
is a narrowband, wireless point-to-point MIMO communication system. The signal vector \( s \in \mathbb{C}^{n_L} \), normalized as \( \mathbb{E}[ss^H] = I \), is to be conveyed over a wireless channel using \( n_T \) transmit antennas and \( n_R \) receive antennas. The channel is characterized in the complex baseband by the channel matrix \( H \in \mathbb{C}^{n_R \times n_T} \). In this setting, the received vector \( y \in \mathbb{C}^{n_R} \) is affinely dependent on the transmitted signal vector \( s \) according to

\[
y = HPs + n,
\]

where \( n \sim \mathcal{C}\mathcal{N}(0, I) \) denotes spatially and temporally white noise. The precoder \( P \in \mathbb{C}^{n_T \times n_L} \) is a design parameter that is used to customize the effective MIMO channel \( HP \in \mathbb{C}^{n_R \times n_L} \). The choice of precoder is subject to a power constraint on the total average transmit power as

\[
\mathbb{E}[||Ps||^2] = \text{tr}(P^H P) \leq P_{\text{max}},
\]

where \( P_{\text{max}} \) is the power budget. We shall limit the number \( n_L \) of simultaneously transmitted symbols over the effective channel by the dimensions of the physical system as \( n_L \leq \min(n_R, n_T) \), which implies that both \( P \) and \( HP \) are tall matrices.

The DF detection procedure can be combined with channel coding \cite{FCG+03}, but is perhaps more easily explained in the setting of uncoded transmission where each \( s_k \) is a symbol in a finite constellation. In contrast to a linear receiver, the DF receiver can exploit this fact by detecting symbols sequentially, in each step aided by feedback from already detected symbols. First, \( s_1 \) is detected as \( \hat{s}_1 = Q(w_1^H y) \), where \( w_1 \) is a suitable equalization vector, and \( Q(\cdot) \) maps a point in the complex plane to the nearest constellation point. It is then possible to use the fact that \( \hat{s}_1 = s_1 \) with high probability in a well-designed system. Prior to detecting \( s_2 \) we may use feedback to mitigate interference caused by transmitting \( s_1 \). By detecting the symbols sequentially we can assume that it is possible to mitigate interference from symbols \( 1, \ldots, k - 1 \) prior to detecting \( s_k \). The procedure can be described in matrix form, as in Figure 3.1, in terms of a feedforward matrix \( W \in \mathbb{C}^{n_R \times n_L} \) and a feedback matrix \( B \in \mathbb{C}^{n_L \times n_L} \). The symbol-vector estimates prior to and after detection, \( \bar{s} \) and \( \hat{s} \) respectively, are related by feedback according to

\[
\bar{s} = W^H y - B\hat{s}.
\]

The feedback matrix \( B \) is assumed to be strictly lower triangular to represent a detection order from \( s_1 \) to \( s_{n_L} \).

### 3.1.1 The Zero-Forcing DF Receiver

Assuming that the receiver knows the effective channel \( HP \), the DF-receiver design problem amounts to optimizing \( W \) and \( B \) based on this
3.1. SYSTEM MODEL

piece of information. These matrices are commonly determined under the simplifying assumption that the feedback is free from error [PJ06,SD08b]. In this case, (3.1) transforms into

\[ \bar{s} = [W^H P B] s + W^H n. \] (3.2)

The receiver design problem can be addressed by minimizing the mean square error (MSE) \( E[|s_k - \bar{s}_k|^2] \) for each \( k = 1, \ldots, n_L \). Employing unconstrained minimum-MSE equalization renders an estimator \( \bar{s} \) of \( s \), which is both biased and admits interference leakage. Another common approach, which is adopted here, minimizes the MSEs under a zero-forcing (ZF) constraint, resulting in an unbiased, interference-free estimator \( \bar{s} \). These ZF properties will prove to be highly useful for the subsequent transmitter design.

We shall first present the ZF-DF receive matrices, and then comment on their optimality. Aided by the QL decomposition \( HP = QL \), where \( Q \in \mathbb{C}^{n_R \times n_L} \) is semi-unitary (satisfying \( Q^H Q = I \)), and \( L \in \mathbb{C}^{n_L \times n_L} \) is lower triangular with strictly positive diagonal entries, the ZF-DF receiver matrices are [PJ06,SD08b]

\[ W_{ZF} = QD_L^{-1}, \quad B_{ZF} = D_L^{-1} L - I, \] (3.3)

where \( D_L \) is a diagonal matrix with the same diagonal as \( L \). The following lemma clarifies the optimality of using the particular ZF matrices in (3.3), which is well-known in the literature. Due to the prominent role of the ZF-DF receiver in this thesis, we prove this result for completeness, and derive the expressions presented in (3.3).

Lemma 3.1. The zero-forcing matrices \( W_{ZF} \) and \( B_{ZF} \) in (3.3) simultaneously minimize all MSEs in (3.2) without trade-off, subject to the following two constraints:

- the ZF constraint \( W^H P B = B + I \);
- the detection order constraint that \( B \) is strictly lower triangular.

Proof. The proof is located in Appendix 3.A.1. \( \square \)

3.1.2 Channel Model and Information

While the receiver is assumed to have access to perfect short-term CSI, we assume that the transmitter only has long-term, statistical CSI. The statistical model of the wireless channel adopted in this chapter is the separable-correlation Rayleigh model as described in detail in Section 2.1.1. It assumes that

\[ H = R_{R}^{1/2} Z R_{T}^{1/2} \] (3.4)
with positive semidefinite transmit correlation $R_T \in \mathbb{C}^{n_T \times n_T}$ and receive correlation $R_R \in \mathbb{C}^{n_R \times n_R}$, and $Z \in \mathbb{C}^{n_R \times n_T}$ has i.i.d. $\mathcal{CN}(0, 1)$ entries. We particularly focus on the case of arbitrary transmit correlation, while the receive antenna array is assumed to be decorrelated. In our model this corresponds to setting $R_R = I$.\(^2\)

### 3.2 Performance Measures

For transmitter optimization it is necessary to discuss how to measure the performance of the ZF-DF transceiver in Rayleigh fading. In the literature, numerous reasonable cost functions appear that reflect system performance [PJ06]. This section analyzes the structure of several important cost functions for statistical precoder design, i.e., the optimization of $P$ based on long-term, statistical channel information. This results in the unifying notion of marginal cost functions. We derive results on marginal cost functions that play an important role in transmitter design in Section 3.3.

#### 3.2.1 Subchannel SNR Characterization

A ZF receiver decomposes the MIMO channel into $n_L$ non-interfering subchannels (assuming correct feedback), and the performance of each of these subchannels can be characterized by its signal-to-noise ratio (SNR). For the ZF signal model

$$\bar{s} = s + W^H n,$$

the SNR of substream $k$ is given by

$$\gamma_k = \frac{\mathbb{E} \left[ |s_k|^2 \right]}{\mathbb{E} \left[ |W_k^H n|^2 \right]} = \frac{1}{w_k^H w_k},$$

where $w_k$ is the $k$th column of the feedforward matrix $W$, using the two assumptions $\mathbb{E} [ss^H] = I$ and $\mathbb{E} [nn^H] = I$. With $W = W_{ZF}$ in (3.3), the SNRs $\gamma_1, \ldots, \gamma_{n_L}$ can be expressed using the Cholesky decomposition $L^H L = P^H H^H H P$ as

$$\gamma_k = l_{kk}^2, \quad k = 1, \ldots, n_L,$$

\(^2\)Note that (3.3) requires an effective channel $HP$ with rank $n_L$. With the channel model we have adopted, this occurs almost surely if and only if $P$ has full column rank, provided that the transmit covariance $R_T$ is positive definite [Mui82]. For the optimal precoders in Section 3.3, $R_T$ need not be positive definite, but $n_L$ should be chosen to be less than or equal to the rank of $R_T$.\(^2\)
3.2. PERFORMANCE MEASURES

where $l_{kk}$ refers to the $k^{\text{th}}$ real and non-negative diagonal element of the lower-triangular matrix $L$. For relations of this kind, we shall make use of the following notation of Cholesky elements.

**Definition 3.1.** Let the positive semidefinite matrix $T \in \mathbb{C}^{N \times N}$ have a Cholesky decomposition $L^H L = T$, where $L \in \mathbb{C}^{N \times N}$ is lower triangular with real, non-negative diagonal elements $l_{11}, \ldots, l_{NN}$. The unique vector of Cholesky elements of $T$ is then defined as $L^2(T) = (l_{11}^2, \ldots, l_{NN}^2)$.

In terms of this notation, we find that the SNR vector is given by

$$\gamma = l^2(P^H H^H H P).$$

(3.7)

Under the transmit-correlated Rayleigh channel model, the matrix $P^H H^H H P \in \mathbb{C}^{n_L \times n_L}$ follows a complex Wishart distribution with $n_R$ degrees of freedom and covariance parameter $P^H R_T P$. The distribution of the Cholesky elements for a Wishart matrix with covariance parameter $I$ is well-known [Goo63, TV04], and the extension to a general covariance parameter is relatively straightforward, which was also realized in [LZW09]. This enables the following SNR characterization, which we pose as a theorem for easy reference. Here and in the following, $\chi^2_{2n}/2$ denotes a scaled chi-squared distribution with probability density function (PDF)$^3$

$$p_n(z) = \frac{z^{n-1} e^{-z}}{(n-1)!}, \quad z \geq 0.$$

**Theorem 3.1.** The subchannel SNRs $\gamma_1, \ldots, \gamma_{n_L}$ for the ZF-DF receiver in transmit-correlated Rayleigh fading are independent and distributed according to

$$\gamma_k \sim x_k \eta_{n_R - n_L + k}, \quad k = 1, \ldots, n_L,$$

where $(x_1, \ldots, x_{n_L}) = l^2(P^H R_T P)$ and each $\eta_n \sim \chi^2_{2n}/2$.

**Proof.** The proof is located in Appendix 3.A.2. \[\square\]

We shall make the subchannel SNRs cornerstones in measuring the performance of a statistically precoded ZF-DF transceiver. All performance measures considered here are based on a mapping from a random SNR vector $\gamma$ into the set of real numbers. We shall denote this mapping by $F_\gamma$. We use this notation to remember that $F_\gamma$ is not a function of $\gamma$, but rather a real-valued functional of the PDF of $\gamma$. By combining a cost functional $F_\gamma$ with the ZF-DF SNR distribution, we obtain a cost function $F_\gamma(x)$, where $x = l^2(P^H R_T P)$.

---

$^3$This distribution is also known as a gamma distribution with shape parameter $n$ and unit scale parameter.
3.2.2 Examples of Performance Measures

Mean Square Error

The communication system at hand decomposes the MIMO channel into $n_L$ scalar subchannels. Regarding $\hat{s}_k$ as an estimator of $s_k$ on each subchannel, the corresponding MSEs $\{f_{\text{MSE}}(\gamma_k)\}_{k=1}^{n_L}$ can be computed from the SNRs $\gamma$. Since the channel is fading, we consider the average MSE with respect to the channel. We may form a cost functional reflecting system performance by combining the average MSEs by weighted summation as

$$F_{\gamma} = \sum_{k=1}^{n_L} \mathbb{E} \left[ \alpha_k f_{\text{MSE}}(\gamma_k) \right]. \quad (3.8)$$

Using uncoded data streams with individual constellations, it is suitable to relate the MSE of subchannel $k$ to the minimum distance $d_{\text{min},k}$ of its constellation by letting $\alpha_k = d_{\text{min},k}^{-2}$. In the special case of square $M$-ary QAM constellations, the minimum distance is given by $d_{\text{min}}^2 = 6/(M-1)$.

It is often desirable to equalize the performance of the subchannels, since the weakest link may have a dominant impact on the overall system performance. This motivates considering a more general cost functional, for example [BPO09]

$$F_{\gamma} = \left\| \left\{ \mathbb{E} \left[ \alpha_k f_{\text{MSE}}(\gamma_k) \right] \right\}_{k=1}^{n_L} \right\|_q, \quad (3.9)$$

where $\|z\|_q = (\sum_k |z_k|^q)^{1/q}$ for $q \geq 1$, and the notation $\{\cdot\}_{k=1}^{n_L}$ refers to an $n_L$-vector. With nonnegative weights, (3.9) reduces to (3.8) when $q = 1$, and in the limiting case as $q \to \infty$,

$$F_{\gamma} = \max_k \left\{ \mathbb{E} \left[ \alpha_k f_{\text{MSE}}(\gamma_k) \right] \right\}_{k=1}^{n_L}.$$

For ZF receivers, MSEs and SNRs are reciprocals by (3.5) and (3.6):

$$f_{\text{MSE}}(\gamma) = \gamma^{-1}. \quad (3.10)$$

Due to the SNR characterization in Theorem 3.1, each subchannel SNR is a random variable on the form $\gamma \sim x\eta_n$, where $\eta_n \sim \chi^2_{2n}/2$. The average MSE under such SNR distributions is easily computed by integration as:\footnote{Note that the average MSE is in fact infinite for $n = 1$, and corresponds to the first data stream in a fully multiplexed system with $n_T \geq n_R = n_L$. This does, however, not mean that the subchannel is useless: Average bit and symbol error probabilities still tend to zero with increasing average SNR.}

$$g_n^{\text{MSE}}(x) = \mathbb{E} \left[ f_{\text{MSE}}(x\eta_n) \right] \quad (3.11)$$

$$= \frac{1}{(n-1)x}, \quad n \geq 2.$$
3.2. PERFORMANCE MEASURES

Joint Symbol Error Probability

The joint symbol error probability (JEP) is an interesting cost function that has been considered before in conjunction with ZF-DF detection [PV04]. The cost functional is simply \( F_\gamma = \Pr(\hat{s} \neq s) \). The instantaneous symbol error probability (SEP) on subchannel \( k \) is denoted by \( f_{k_{\text{SEP}}}^{\gamma_k} \). Due to the independence of the SNRs in Theorem 3.1, and the diagonal noise covariance with \( W = W_{ZF} \) in (3.5), it follows that the JEP is given by

\[
F_\gamma = 1 - \prod_k \left[ 1 - \mathbb{E}\left[f_{k_{\text{SEP}}}^{\gamma_k}\right]\right].
\]  

(3.12)

The ZF-DF receiver establishes non-interfering additive white Gaussian noise (AWGN) subchannels. When a square \( M \)-ary QAM constellation is used on an AWGN subchannel, the SEP is given by [Pro00]

\[
f_{k_{\text{SEP}}}^{\gamma_k} = 2\theta \text{erfc} \left( \sqrt{\beta \gamma} \right) - \theta^2 \text{erfc}^2 \left( \sqrt{\beta \gamma} \right),
\]

where \( \theta = (\sqrt{M} - 1)/\sqrt{M} \), \( \beta = 3/(2M - 2) \), and \( \text{erfc}(\cdot) \) denotes the complementary error function. The corresponding average SEP with \( \gamma \sim x\eta_n \) and \( \eta_n \sim \chi_{2n}/2 \) is then given by

\[
g_{n_{\text{SEP}}}^{\gamma_n}(x) = \mathbb{E}\left[f_{n_{\text{SEP}}}^{\gamma_n}(x\eta_n)\right]
= 2\theta g_{n_{\text{erfc}}}^{\gamma_n}(\beta x) - \theta^2 g_{n_{\text{erfc}}}^{\gamma_n}(\beta x),
\]

(3.14)

in terms of the two auxiliary functions \( g_{n_{\text{erfc}}}^{\gamma_n}(x) = \mathbb{E}\left[\text{erfc}\left(\sqrt{x\eta_n}\right)\right] \) and \( g_{n_{\text{erfc}}}^{\gamma_n^2}(x) = \mathbb{E}\left[\text{erfc}^2\left(\sqrt{x\eta_n}\right)\right] \). These two expectations are known [SA04] and can be expressed as

\[
g_{n_{\text{erfc}}}^{\gamma_n}(x) = 1 - \sqrt{\frac{x}{1 + x}} \sum_{m=0}^{n-1} T_{nm}(1 + x)^{-m},
\]

(3.15)

\[
g_{n_{\text{erfc}}}^{\gamma_n^2}(x) = 1 - (1 - g_{n_{\text{erfc}}}^{\gamma_n}(x)) \frac{4}{\pi} \arctan \sqrt{\frac{1 + x}{x}} + \frac{4}{\pi} \sum_{m=1}^{n-1} \sum_{i=1}^{m} T_{im} \frac{x}{[1 + x]^i[1 + 2x]^{m-i+1}},
\]

using constants \( T_{im} \) defined as

\[
T_{im} = \begin{pmatrix} \binom{2m}{m} \\ \binom{2(m-i)}{m-i} \end{pmatrix} 4^i [2(m - i) + 1] \]

Note that although \( f_k^{\text{SEP}}(\gamma_k) \) is the SEP conditioned on correct feedback, (3.12) is the true JEP.
CHAPTER 3. FRAMEWORK FOR PRECODER OPTIMIZATION

Bit Error Probability

The MIMO system under consideration can be utilized to demultiplex a single bitstream onto the \( n_L \) subchannels. We assume that for each subchannel \( k \), Gray coding is used to map bits into complex symbols in a square \( M_k \)-ary QAM constellation. Then the bit error probability (BEP) on each subchannel, \( f_{\text{BEP}}^{k}(\gamma_k) \), can be computed. With \( s_k \) representing \( \log_2 M_k \) bits, the average system BEP (assuming no error propagation) is given by

\[
\mathcal{F}_\gamma = \frac{1}{\sum m \log_2 M_m} \sum_{k=1}^{n_L} \mathbb{E} \left[ f_{\text{BEP}}^{k}(\gamma_k) \right] \log_2 M_k. \tag{3.16}
\]

For a complex AWGN subchannel and a square \( M \)-ary QAM constellation, the BEP is related to the SNR \( \gamma \) according to \[CY02\]

\[
f_{\text{BEP}}(\gamma) = \sum_{m=0}^{\sqrt{M}-2} \delta_m \text{erfc} \left( (2m + 1)\sqrt{\beta \gamma} \right). \tag{3.17}
\]

As before, \( \beta = 3/(2M - 2) \), and the constants \( \delta_m \) are given by

\[
\delta_m = \sum_{i=\left\lfloor \log_2 \sqrt{M} \right\rfloor}^{\left\lfloor \log_2 \sqrt{M} + 1 \right\rfloor} (-1)^{\left\lfloor m \cdot 2^i - 1 \right\rfloor} \sqrt{M} \log_2 \sqrt{M} \left( 2^{i-1} - \left\lfloor m \cdot 2^i - 1 \right\rfloor \sqrt{M} + 1 \right) \tag{3.18}.
\]

The average BEP with \( \gamma \sim x \eta_n \) and \( \eta_n \sim \chi^2_{2n}/2 \) similarly becomes

\[
g_{\text{BEP}}^{n}(x) = \mathbb{E} \left[ f_{\text{BEP}}(x \eta_n) \right] \tag{3.19}
= \sum_{m=0}^{\sqrt{M}-2} \delta_m g_{\text{erfc}}^{n} ((2m + 1)^2 \beta x),
\]

with \( g_{\text{erfc}}^{n}(x) \) given by (3.15).

Outage Formulation

Assume that it is only possible to maintain reliable communication on subchannel \( k \) if its SNR \( \gamma \) is above a certain threshold \( \gamma_k^{\text{th}} \). Then it is of interest to consider the probability of outage, \( \Pr \left( \gamma \leq \gamma_k^{\text{th}} \right) \). When several parallel subchannels are present, a suitable cost functional is

\[
\mathcal{F}_\gamma = \Pr \left( \min_k \frac{\gamma_k}{\gamma_k^{\text{th}}} \leq 1 \right), \tag{3.20}
\]
which declares an outage if any subchannel is in outage. Assuming that the components of $\gamma$ are independent random variables, which by Theorem 3.1 is the case considered here, it is possible to reformulate (3.20) as

$$F_\gamma = 1 - \prod_k \left[ 1 - \Pr(\gamma_k \leq \gamma_{th}^k) \right]. \quad (3.21)$$

The probability of outage on a subchannel with $\gamma \sim x\eta_n$ and $\eta_n \sim \chi^2_{2n}/2$ is given by

$$g_{Outage}^n (x; \gamma_{th}) = \Pr(x\eta_n \leq \gamma_{th}) = 1 - e^{-\gamma_{th}/x} \sum_{m=0}^{n-1} \frac{\left(\frac{\gamma_{th}}{x}\right)^m}{m^m}. \quad (3.21)$$

For the following discussion in Section 3.2.3, we note that the outage probability can also be regarded as an average cost function. It is possible to choose $f_{Outage}^\gamma (\gamma)$ such that

$$\Pr(\gamma \leq \gamma_{th}) = \mathbb{E} \left[ f_{Outage}^\gamma (\gamma) \right]$$

by assigning a unit cost only for SNRs $\gamma \leq \gamma_{th}$,

$$f_{Outage}^\gamma (\gamma) = \begin{cases} 1, & \text{if } 0 \leq \gamma \leq \gamma_{th}, \\ 0, & \text{if } \gamma > \gamma_{th}. \end{cases} \quad (3.22)$$

**Achievable Data Rate**

The maximal data rate that can be reliably maintained in fast fading, subject to ZF-DF receive processing and $n_L$ independently encoded datastreams, is given by adding the average Gaussian mutual information on the subchannels [TV05, Sec. 8.3]. Following the convention of the other examples that $F_\gamma$ is a cost to be minimized, the negative achievable data rate (in nats) is given by

$$F_\gamma = \sum_{k=1}^{n_L} \mathbb{E} \left[ f_{Rate} (\gamma_k) \right], \quad (3.23)$$

where

$$f_{Rate} (\gamma) = -\log(1 + \gamma). \quad (3.24)$$

The average negative mutual information with $\gamma \sim x\eta_n$ and $\eta_n \sim \chi^2_{2n}/2$ is given by [SA04] as

$$g_{Rate}^n (x) = \mathbb{E} \left[ f_{Rate} (x\eta_n) \right] = -\mathcal{P}_n (-x^{-1}) E_1(x^{-1}) - \sum_{m=1}^{n-1} \frac{\mathcal{P}_m (x^{-1}) \mathcal{P}_{n-m} (-x^{-1})}{m} \quad (3.25)$$
where $E_1(z) = \int_1^{\infty} e^{-zt/t} \, dt$, and $P_m(z) = e^{-z} \sum_{j=0}^{m-1} z^j/j!$.

### 3.2.3 Properties of Performance Measures

By reviewing the example performance measures (3.9), (3.12), (3.16), (3.21), and (3.23), we find that these have a certain common structure, and are based solely on the *marginal* distributions of the SNR vector $\gamma$. This structure is embodied in the following definition of a *marginal cost functional*.

**Definition 3.2.** A marginal cost functional is on the form

$$F_\gamma = G\left(\left\{ \mathbb{E} \left[ f_k(\gamma_k) \right] \right\}_{k=1}^{n_L} \right),$$

where

- $f_k : \mathbb{R}_+ \to \mathbb{R}$ is decreasing for $k = 1, \ldots, n_L$, and
- $G : \mathbb{R}^{n_L} \to \mathbb{R}$ is increasing in each argument.

A marginal cost functional $F_\gamma$ determines the total system cost in two separate steps. First, on each subchannel $k$, the *inner cost function* $f_k : \mathbb{R}_+ \to \mathbb{R}$ specifies the cost $f_k(\gamma_k)$ associated with having SNR $\gamma_k$ on this subchannel. The average cost is then computed in order to ensure the dependence on the PDF of $\gamma_k$. Second, an *outer cost function* $G : \mathbb{R}^{n_L} \to \mathbb{R}$ combines the average costs into the overall cost of the system.

By combining a cost functional $F_\gamma$ with the ZF-DF SNR distribution, we obtain a *cost function* $F_\gamma(x)$, where $x = l^2(P^H R T P)$. For a *marginal cost function* induced by Definition 3.2, we may denote the average cost of the $k$th substream as

$$g_k(x_k) = \mathbb{E} \left[ f_k(\gamma_k) \right], \quad (3.26)$$

which results in

$$F_\gamma(x) = G\left(\{g_k(x_k)\}_{k=1}^{n_L}\right). \quad (3.27)$$

We shall state two basic results on marginal cost functions. Any marginal cost function is componentwise decreasing, and convex if the outer cost function as well as all the expected inner cost functions are convex. In the following we shall make use of this fact in logarithmic variables $\bar{x} = \log(x)$, and we therefore present these results accordingly.

**Lemma 3.2.** Any cost function $F_\gamma(e^{\bar{x}})$ based on a marginal cost functional $F_\gamma$, with $\gamma$ and $\bar{x} = \log(x)$ from Theorem 3.1, has the following properties:

- $F_\gamma(e^{\bar{x}})$ is decreasing in each component of $\bar{x}$;
3.2. PERFORMANCE MEASURES

- $\mathcal{F}_{\gamma}(e^{\bar{x}})$ is convex in $\bar{x}$ if $\mathcal{G}$ is convex and every $g^k(e^{\bar{x}})$ is convex in $\bar{x}$ for $k = 1, \ldots, n_L$.

Proof. It follows from (3.26), and each $f^k(x)$ being decreasing in $x$, that each $g^k(e^{\bar{x}})$ is decreasing in $\bar{x}$. The first result then follows by $\mathcal{G}$ being increasing in each argument. The second result is concluded by applying a standard convexity-preserving composition rule [BV04], using the fact that $\mathcal{G}$ is both convex and increasing.

We have not yet commented on the convexity of average inner cost functions on the form $g(e^{\bar{x}}) = \mathbb{E}[f(e^{\bar{x}}\eta_n)]$ for $\eta_n \sim \chi^2_{2n/2}$. General results on convexity are hard to find. By examining the examples considered previously, we find that convexity holds for several performance measures for function values of interest. Indeed, the average MSE $g^{\text{MSE}}$ is convex in $\bar{x}$ for any $\bar{x} \in \mathbb{R}$, and the outage probability $g^{\text{Outage}}$ is convex for values below $0.5$.\footnote{\textsuperscript{6}} For symbol and bit error probabilities with square QAM constellations, convexity holds for $g^{\text{SEP}} \leq 0.2$ and $g^{\text{BEP}} \leq 0.1$, respectively.\footnote{\textsuperscript{7}} Interestingly, the negative average mutual information $g^{\text{Rate}}$ turns out to be concave in $\bar{x}$, and will be given special attention in Section 3.3.3.

In order to further analyze the cost functions, we shall introduce the notion of majorization [MOA11], which has shown to be tremendously useful in previous work on MIMO transceiver design [PCL03, SD08a, JPV07]. A common definition of majorization is the following:

\textit{Definition 3.3.} For $a, b \in \mathbb{R}^N$, $a$ is said to be majorized by $b$, or $a \preceq b$, if and only if

\begin{align*}
\sum_{k=1}^{n} a_{\downarrow k} &\leq \sum_{k=1}^{n} b_{\downarrow k}, \quad n = 1, \ldots, N - 1, \\
\sum_{k=1}^{N} a_k &\leq \sum_{k=1}^{N} b_k,
\end{align*}

where $a_{\downarrow k}$ and $b_{\downarrow k}$ denote the $k^{\text{th}}$ largest component of $a$ and $b$, respectively.

The majorization relation may be regarded as a partial order among vectors. The real-valued functions that preserve this order are intricately related to this concept. These are termed Schur-convex/Schur-concave and are defined as follows.

\textsuperscript{6}It can be shown by differentiation that $g^{\text{Outage}}_{\gamma}(e^{\bar{x}}; \gamma^{\text{th}})$ is convex on $\bar{x} \geq \log(\gamma^{\text{th}}/n)$. By examining the function values $g^{\text{Outage}}_{\gamma}(\gamma^{\text{th}}/n; \gamma^{\text{th}})$ for different $n$ numerically, it is possible to conclude that convexity holds in general for outage probabilities $g^{\text{Outage}} \leq 0.5$.

\textsuperscript{7}We have verified this numerically for SNR distributions $\gamma \sim e^{\bar{x}}\eta_n$ with $\eta_n \sim \chi^2_{2n/2}$ for $n \leq 20$ and QAM-modulations up to $M = 4096$. 
Definition 3.4. A real-valued function \( \phi : \mathbb{R}^N \to \mathbb{R} \) is said to be Schur-convex if \( \mathbf{a} \preceq \mathbf{b} \) implies \( \phi(\mathbf{a}) \leq \phi(\mathbf{b}) \). If, on the other hand, \( \mathbf{a} \preceq \mathbf{b} \) implies \( \phi(\mathbf{a}) \geq \phi(\mathbf{b}) \), then \( \phi \) is said to be Schur-concave.

By reviewing the previous list of performance measures, we find several examples of outer cost functions \( G \) that are Schur-convex or Schur-concave on \( \mathbb{R}^n \). For example,

- \( G(\mathbf{z}) = (\sum_k |z_k|^q)^{1/q} \) is Schur-convex for \( q \geq 1 \),
- \( G(\mathbf{z}) = 1 - \prod_k (1 - z_k) \) is Schur-convex,
- \( G(\mathbf{z}) = \sum_k z_k \) is both Schur-convex and Schur-concave.

The first example appears in the MSE formulation (3.8) with identity weights, while the second example applies to the JEP and the outage formulation in (3.12) and (3.21), respectively. The third example is used to compose the achievable data rate in (3.23), and demonstrates that a function can be both Schur-convex and Schur-concave. Important subclasses of Schur-convex and Schur-concave functions are those that are convex and concave, respectively, in addition to being symmetric\(^8,9\).

In section 3.3.1 we devote special attention to marginal cost functionals that are also symmetric in the SNRs:

Definition 3.5. A marginal cost functional in Definition 3.2 is called symmetric, and is denoted by \( F(\gamma) \), if it satisfies

- \( f_1 = f_2 = \ldots = f_n \), and
- \( G \) is a symmetric function.

Although the cost functional \( F(\gamma) \) is symmetric in \( \gamma \), the induced cost function \( F(\gamma)(\mathbf{x}) \) in (3.27) is certainly not symmetric in \( \mathbf{x} \). Nevertheless, as we show next, important properties of such cost functions can still be extracted in a general setting. The following theorem uses the notation that \( \bar{x}_\downarrow \) and \( \bar{x}_\uparrow \) are the vectors formed by rearranging the components of \( \bar{x} \) into decreasing and increasing order, respectively.

Theorem 3.2. Assume that

\[
F(\gamma)(\mathbf{x}) = G \left( \{g^k(x_k)\}_{k=1}^{n_L} \right)
\]

\(^8\)A function \( \phi \) is symmetric if \( \phi(\mathbf{z}) = \phi(\Pi \mathbf{z}) \) for any \( \mathbf{z} \in \mathbb{R}^N \) and any \( N \)-dimensional permutation matrix \( \Pi \).

\(^9\)Note that \( G(\mathbf{z}) = 1 - \prod_k (1 - z_k) \) is Schur-convex, but in fact not convex. However, for precoder optimization to be addressed in the following, we may equivalently minimize \( 1 - \log(1 - G) \), which is symmetric and convex.
3.3. PRECODER OPTIMIZATION

is given by combining a symmetric marginal cost functional \( F^S_\gamma \) with the ZF-DF SNR distribution in Theorem 3.1. If \( G \) is Schur-convex, and each \( g^k(e^{\bar{x}}) \) is convex in \( \bar{x} \), then

\[
F^S_\gamma (e^{\bar{x}_i}) \leq F^S_\gamma (e^{\bar{x}}), \quad \forall \bar{x} \in \mathbb{R}^{n_L}.
\] (3.28)

If, on the other hand, \( G \) is Schur-concave, and each \( g^k(e^{\bar{x}}) \) is concave in \( \bar{x} \), then

\[
F^S_\gamma (e^{\bar{x}_i}) \leq F^S_\gamma (e^{\bar{x}}), \quad \forall \bar{x} \in \mathbb{R}^{n_L},
\] (3.29)

and \( F^S_\gamma (e^{\bar{x}}) \) is also Schur-concave in \( \bar{x} \) on the set of decreasing vectors \( \{ \bar{x} \in \mathbb{R}^{n_L} : \bar{x}_1 \geq \ldots \geq \bar{x}_{n_L} \} \).

Proof. A proof is provided in Appendix 3.A.3. \( \square \)

According to the first part of Theorem 3.2, all convex cost functions \( F^S_\gamma (e^{\bar{x}}) \) based on a symmetric marginal cost functional behave similarly under optimization with respect to \( \bar{x} \); minimizing \( F^S_\gamma (e^{\bar{x}}) \) over all permutations of a given \( \bar{x} \) results in the minimizer \( \bar{x}_1 \), which is in opposite order to the substream decoding order 1, \ldots, n_L. The optimal order stems from the different statistical properties of the SNRs in Theorem 3.1. The result naturally extends to any other fixed decoding order; it is optimal to let \( \bar{x} \) be arranged in the opposite order. In the terminology of [MOA11, Sec. 6.F], we have shown that the entire class of convex cost functions based on a symmetric marginal cost functional is “arrangement increasing” in the decoding order and \( \bar{x} \). The second part of the theorem gives conditions under which the cost function is “arrangement decreasing” in the decoding order and \( \bar{x} \).

3.3 Precoder Optimization

With a ZF-DF receiver, system performance depends on the SNR vector \( \gamma \), which is distributed according to Theorem 3.1. The set of possible distributions is parameterized by \( \bar{x} = l^2(P^H R_T P) \). We consider the transmitter design problem of finding a suitable precoder \( P \) assuming a constraint on the total average transmit power, \( \text{tr} (P^H P) \leq P_{\text{max}} \). An optimization problem designed to maximize system performance can therefore take the form in

\textbf{Design Formulation 3.1.}

\[
\begin{align*}
\text{minimize} & \quad F_\gamma (\bar{x}) \\
\text{subject to} & \quad \bar{x} = l^2(P^H R_T P), \\
& \quad \text{tr} (P^H P) \leq P_{\text{max}}.
\end{align*}
\]
Our analysis of the transmitter optimization problem is twofold. In the first part, we use recent advances in MIMO transceiver design to convert Design Formulation 3.1 into a convex optimization problem. We propose a novel formulation that can be readily implemented and solved using, for example, a primal-dual interior-point method [BV04]. In the second part, we focus on certain classes of performance measures that admit a reduced-complexity formulation for precoder optimization.

We shall address Design Formulation 3.1 in the structured manner outlined in, for example, [PJ06, VP07]: A useful tool in finding the optimal precoder \( P \in \mathbb{C}^{n_T \times n_L} \) is the singular-value decomposition (SVD). Without loss of generality, we may assume that the precoder is synthesized as \( P = UD \left( \sqrt{p} \right) V_H \), with a beamforming matrix \( U \in \mathbb{C}^{n_T \times n_L} \) having orthonormal columns, a power allocation \( p \in \mathbb{R}^{n_L} \) with positive components, and a unitary input-shaping matrix \( V \in \mathbb{C}^{n_L \times n_L} \). Here, \( D \left( \sqrt{p} \right) \in \mathbb{R}^{n_L \times n_L} \) refers to the diagonal matrix having \( \sqrt{p} = (\sqrt{p_1}, \ldots, \sqrt{p_{n_L}}) \) on the diagonal. Our first goal is to convert Design Formulation 3.1 into an optimization problem with a convex domain. The next theorem states the optimality of choosing a beamforming matrix \( U \) that matches the \( n_L \) statistically strongest directions of the channel. It requires the eigenvalue decomposition of the transmit correlation matrix \( R_T = U_T D (\lambda_T) U_T^H \), where \( \lambda_T \in \mathbb{R}^{n_T} \) is the decreasing vector of eigenvalues, and \( U_T = [u_{T,1} \ldots u_{T,n_T}] \) is the unitary matrix of corresponding eigenvectors.

**Theorem 3.3.** Assuming that \( \mathcal{F}_\gamma (x) \) is componentwise decreasing in \( x \), there is a minimizer \( P = UD \left( \sqrt{p} \right) V_H \) of Design Formulation 3.1 with beamforming matrix

\[
U = [u_{T,1} \ldots u_{T,n_L}].
\]

**Proof.** A proof is provided in Appendix 3.A.4. \( \Box \)

Transmitting in the \( n_L \) statistically strongest directions of the channel, as suggested by Theorem 3.3, results in

\[
x = l^2 (P^H R_T P) = l^2 (VD \left( p \odot \lambda_T^l \right) V_H),
\]

where \( \odot \) denotes componentwise (Schur) multiplication, and \( \lambda_T^l \) is a decreasing vector of the \( n_L \) largest eigenvalues of \( R_T \). The next step is to identify the set of vectors \( x \) that is generated by varying the input-shaping matrix \( V \). The answer has already appeared as a lemma in the proof of Theorem 3.3, and we state it again for clarity.

**Lemma 3.3.** Given two \( n_L \)-vectors \( x \) and \( z \) with real, positive components, there is a unitary matrix \( V \in \mathbb{C}^{n_L \times n_L} \) such that \( x = l^2 (VD (z) V^H) \) if and only if \( \log(x) \preceq \log(z) \).

\(^{10}\)The logarithm is taken componentwise.
3.3. PRECODER OPTIMIZATION

Finding the unitary matrix $V$ that establishes $x = l^2(VD(z)V^H)$ for a given pair $(x, z)$ in Lemma 3.3 is a straightforward application of the generalized triangular decomposition [JHL08]. The usefulness of Lemma 3.3 is the following. Instead of optimizing $V$ over the Stiefel manifold of unitary $n_L \times n_L$ matrices, we may instead optimize $x$ over the domain specified by $\log(x) \preceq \log(p \odot \lambda^T_l)$. Assuming a full column-rank $P$, which is required for the ZF receiver, we may switch into logarithmic variables $\bar{x} = \log x$ and $\bar{p} = \log p$. If, similarly, $\bar{\lambda}_T^l = \log \lambda^T_T$ the transformed optimization problem becomes

\begin{align*}
\text{minimize} \quad & \mathcal{F}_\gamma(e^\bar{x}) \\
\text{subject to} \quad & \bar{x} \preceq \bar{p} + \bar{\lambda}_T^l, \\
& \sum_{k=1}^{n_L} e^{\bar{p}_k} \leq P_{\text{max}}.
\end{align*}

Optimization problems with majorization constraints occur frequently in the literature on MIMO transceiver design. Almost exclusively, it is assumed that the objective $\mathcal{F}_\gamma(e^\bar{x})$ is minimized when the components of $\bar{x}$ are arranged in a particular order, which can be figured out in advance. While this may be a reasonable assumption in several contexts, it is not the general case studied here. Indeed, with different coding and modulation schemes on the subchannels, in combination with the different fading statistics of the subchannels established in Theorem 3.1 the optimal order is not known in general.

While considerably less attention has been given to the case of a general objective in the literature, the necessary steps towards an optimization problem with a convex domain are provided in [PJ06]: First, we may assume that $\bar{p}$ has elements in decreasing order. Indeed, if $(\bar{x}, \bar{p})$ is a global minimizer of Design Formulation 3.2 then so is $(\bar{x}, \bar{p}_\downarrow)$, where $\bar{p}_\downarrow$ is the rearrangement of $\bar{p}$ in decreasing order, since $\bar{p} + \bar{\lambda}_T^l \preceq \bar{p}_\downarrow + \bar{\lambda}_T^l$. Second, by the very definition of majorization in Definition 3.3 we end up with the following optimization problem.

\footnote{While originally developed for MIMO transceiver design, this general matrix decomposition has found a variety of applications: see, e.g., [Wen11] for a recent Ph.D thesis on the topic.}
**Design Formulation 3.3.**

\[
\begin{align*}
\text{minimize} & \quad \mathcal{F}_\gamma (e^\bar{x}) \\
\text{subject to} & \quad \sum_{k=1}^n \bar{x}_{\downarrow k} \leq \sum_{k=1}^n \bar{p}_k + \bar{\lambda}_{T,k}, \quad n = 1, \ldots, n_L - 1, \\
& \quad \sum_{k=1}^{n_L} \bar{x}_k = \sum_{k=1}^{n_L} \bar{p}_k + \bar{\lambda}_{T,k}, \\
& \quad \sum_{k=1}^{n_L} e^{\bar{p}_k} \leq P_{\text{max}}.
\end{align*}
\]

Note that we have omitted the constraint that \(\bar{p}\) should be decreasing, since this constraint becomes superfluous in the new context. In this formulation the domain is convex in the variables \((\bar{x}, \bar{p})\), since each function \(\sum_{k=1}^n \bar{x}_{\downarrow k}\) is convex in \(\bar{x}\).

The main drawback of Design Formulation 3.3 is that constraints involving the function \(\sum_{k=1}^n \bar{x}_{\downarrow k}\) are non-differentiable. A solution can still be obtained using subgradient-based algorithms for constrained optimization [Ber99], but the convergence of such algorithms is known to be quite slow. In comparison, primal-dual interior point methods have much faster convergence, but require twice differentiable constraints and objective [BV04].

As we show next, it is possible to implement the majorization constraint in an efficient manner without compromising neither differentiability nor the generality of the objective. It relies on the key observation

\[
\sum_{k=1}^n \bar{x}_{\downarrow k} = \min_{u_n, t_n} \left\{ 1^T u_n + n t_n : 1 t_n + u_n \geq \bar{x}, u_n \geq 0 \right\},
\]

where \(0\) and \(1\) denote \(n_L\)-vectors with only zeroes and ones, respectively. The correctness of (3.30) can be verified using duality theory of linear programs as suggested in [BV04, p. 278]. For a more direct approach, it is easy to conclude that \(\sum_{k=1}^n \bar{x}_{\downarrow k} \leq 1^T u_n + n t_n\) for any feasible pair \((u_n, t_n)\), and equality is attained for any \(t_n\) in the interval \(\bar{x}_{\downarrow (n+1)} \leq t_n \leq \bar{x}_{\downarrow n}\) and associated \(u_n = (u_{n,1}, \ldots, u_{n,n_L})\) with \(u_{n,k} = \max(0, \bar{x}_k - t_n)\).

---

12As mentioned in [PCL03], it is possible to optimize directly with respect to the decreasing vector \(\bar{x}_{\downarrow}\), which makes the constraints affine, and hence differentiable, in \((\bar{x}_{\downarrow}, \bar{p})\). However, this comes at the price of swapping \(\mathcal{F}_\gamma (e^\bar{x})\) for the modified objective \(\min_{\Pi} \mathcal{F}_\gamma (e^{\Pi^\top \bar{x}})\), which finds the optimal order by exhaustive search over the set of \(n_L \times n_L\) permutation matrices \(\Pi\). This is not a remedy to the problem, since the modified objective is in general non-differentiable. Moreover, this approach is awkward from a complexity point of view, since the number of permutation matrices of size \(n_L \times n_L\) scales like \(n_L!\).
3.3. PRECODER OPTIMIZATION

We shall apply the result (3.30) for each constraint in Design Formulation 3.3 that involves \( \sum_{k=1}^{n} \bar{x}_{nk} \) for \( n = 2, \ldots, n_L - 2 \). Introducing auxiliary variables \( u_n \) and \( t_n \) for each \( n \), with corresponding feasibility constraints \( 1t_n + u_n \geq \bar{x} \) and \( u_n \geq 0 \), enables the following implementation with \( 2n_L^2 - 3n_L - 1 \) constraints and \( n_L^2 - 3 \) variables.

**Design Formulation 3.4.**

\[
\begin{align*}
\text{minimize} & \quad \mathcal{F}_\gamma (\bar{e} \bar{x}) \\
\text{subject to} & \quad \bar{p}_{nL} + \lambda_{T,nL} + \bar{x}_k \leq \bar{p}_1 + \lambda_{T,1}, \quad k = 1, \ldots, n_L, \\
& \quad 1^T u_n + n t_n \leq \sum_{k=1}^{n} \bar{p}_k + \lambda_{T,k}, \\
& \quad 1^T u_n \geq \bar{x}, \quad n = 2, \ldots, n_L - 2, \\
& \quad u_n \geq 0, \\
& \quad \sum_{k=1}^{n_L} x_k = \sum_{k=1}^{n_L} \bar{p}_k + \lambda_{T,k}, \\
& \quad \sum_{k=1}^{n_L} e^{\bar{p}_k} \leq P_{\text{max}}.
\end{align*}
\]

In Design Formulation 3.4, all constraints are linear in the optimization variables except for the last constraint, which involves a convex, smooth function in \( \bar{p} \). Provided that the objective function is convex and twice differentiable, this is a convex optimization problem in standard form that can be very efficiently implemented. From our list of example cost functions, Design Formulation 3.4 applies to the minimization of MSE-based costs, as well as the JEP, BEP, and probability of outage.

Design Formulation 3.4 also reveals an interesting feature on the implementation of majorization constraints in optimization. In general, the majorization constraint embodies an \((n_L - 1)\)-dimensional polyhedron in \( \mathbb{R}^{n_L} \) with \( n_L! \) vertices and \( 2^{n_L} \) faces. Still, such a constraint can be implemented with a quadratic number of variables and affine constraints. As demonstrated in Design Formulation 3.4, this holds even in the quite intricate setting of Design Formulation 3.2, where the majorization constraint is coupled with the transmit-power constraint on \( \bar{p} \).

The number of constraints and variables can be reduced further, growing linearly in \( n_L \), if the optimal order of the vector \( \bar{x} \) can be figured out in advance. This turns out to be the case for several important classes of cost functions, which we investigate next.

### 3.3.1 Equal-Rate Designs

Many works on MIMO transceiver design in the literature propose using the same data rate on each subchannel, for example [KS04b, JLH05, LZW09]. In
such a case, it is often desirable to equalize the subchannel SNRs. With perfect CSI available at the transmitter (i.e., the channel matrix $H$ is known), it is indeed possible to ensure that $\gamma = l^2(P^H H^H H P)$ becomes a uniform vector using the geometric-mean decomposition [JLH05]. This approach cannot be applied in the setup considered here with only statistical channel information at the transmitter. Equalizing the power-allocation vector $x = l^2(P^H R_T P)$ still gives subchannel SNR distributions that are highly unsymmetric. It is easily verified using Theorem 3.1 that, for any $\gamma_{th} > 0$,

$$\Pr (\gamma_{nL} \geq \gamma_{th}) \geq \ldots \geq \Pr (\gamma_1 \geq \gamma_{th}),$$

which concludes in a strong sense that subchannels placed later in the decoding order will have better performance. From the point of view of error propagation the reverse situation would be preferable. In order to mitigate this effect it is necessary to let $x$ be decreasing in its components. Enforcing such a solution in Design Formulation 3.3 leads to the following simplified optimization problem.

**Design Formulation 3.5.**

$$\begin{align*}
\text{minimize} & \quad F_{\gamma} (e^x) \\
\text{subject to} & \quad \bar{x}_n \geq \bar{x}_{n+1}, \quad n = 1, \ldots, n_L - 1, \\
& \quad \sum_{k=1}^{n} \bar{x}_k \leq \sum_{k=1}^{n} \bar{p}_k + \bar{\lambda}_{T,k}, \quad n = 1, \ldots, n_L - 1, \\
& \quad \sum_{k=1}^{n_L} \bar{x}_k = \sum_{k=1}^{n_L} \bar{p}_k + \bar{\lambda}_{T,k}, \\
& \quad \sum_{k=1}^{n_L} e^\bar{p}_k \leq P_{max}.
\end{align*}$$

Optimizing over decreasing vectors $\bar{x}$ is both intuitive and leads to an efficiently implementable design formulation. It is, however, also possible to state the optimality of this approach by analyzing how reasonable cost functions $F_{\gamma} (x)$ behave under optimization with respect to rearrangements of the components of $x$. In the case of an equal-rate design, we may assume a certain symmetry in the way we measure performance of the system. For a marginal cost function, the same cost $f (\gamma)$ should be assigned to any subchannel with instantaneous SNR $\gamma$, and the expected costs $E [f (\gamma_k)]$ should be combined into the overall system cost in a symmetric manner by letting the outer cost function $G$ be symmetric in its arguments. This amounts to the definition of a symmetric marginal cost functional in Definition 3.5, for which the following holds.
3.3. PRECODER OPTIMIZATION

Theorem 3.4. Assume that $\mathcal{F}_\gamma(x) = \mathcal{F}^S_\gamma(x)$ is based on a symmetric marginal cost functional, with convex expected costs $g^k(e^x)$. Then any minimizer of Design Formulation 3.5 is also a minimizer of Design Formulation 3.2.

Proof. Let $(\bar{x}, \bar{p})$ be any feasible point in Design Formulation 3.2. Then $(\bar{x}^\downarrow, \bar{p})$, where $\bar{x}^\downarrow$ is the rearrangement of $\bar{x}$ in decreasing order, is feasible in both Design Formulation 3.2 and 3.5. Theorem 3.2 ensures that $\mathcal{F}^S_\gamma(e^{\bar{x}}) \geq \mathcal{F}^S_\gamma(e^{\bar{x}^\downarrow})$. Moreover, $\mathcal{F}^S_\gamma(e^{\bar{x}^\downarrow}) \geq \mathcal{F}^S_\gamma(e^{\bar{x}^\downarrow})$ for any minimizer $(\bar{x}^*_\downarrow, \bar{p}^*)$ of Design Formulation 3.5. Since the domain of Design Formulation 3.5 is a subset of the domain of Design Formulation 3.2, the result follows.

The key observation resulting in Theorem 3.4 is the behavior of convex cost functions $\mathcal{F}^S_\gamma(e^{\bar{x}})$ under optimization over permutations, which was studied in Theorem 3.2; for any $\bar{x}$ it holds that $\mathcal{F}^S_\gamma(e^{\bar{x}^\downarrow}) \leq \mathcal{F}^S_\gamma(e^{\bar{x}})$. This property was identified for the expected sum-MSE cost function in [LZW09], which considers a similar precoding problem under the dimensional constraints $n_R > n_T = n_L$. Theorem 3.2 reveals that the arrangement property holds for all convex cost functions based on a symmetric marginal cost functional. While being quite easily shown for the expected sum-MSE, it is not trivial to prove the more general result in Theorem 3.2. This enables Theorem 3.4 to generalize the results in [LZW09] to regard many relevant cost functions, in addition to relaxing the constraints on the dimensions to $n_L \leq \min(n_R, n_T)$. In particular, the result holds for any MSE-based cost function with equal weights, for bit and joint symbol error probabilities with equal-rate subchannels using QAM constellations, as well as for the outage probability $\Pr(\min_k \gamma_k \leq \gamma_{th})$.

3.3.2 MSE-based Designs

We shall return to the more general setting allowing non-uniform rate allocations. The symmetric cost function of Definition 3.5 is in general inadequate, since the cost of having a certain SNR on a subchannel should depend on the associated rate. As we saw in the case of equal-rate designs, Design Formulation 3.3 can be simplified if the optimal order of $\bar{x}$ can be figured out in advance. This section deals with certain MSE-based cost functions, for which it is possible to determine the optimal order.

We consider the class of MSE-based cost functions in (3.9),

$$\mathcal{F}_\gamma = \| \{ \mathbb{E} [\alpha_k f^{MSE}(\gamma_k)] \}_{k=1}^{n_L} \|_q,$$
where the average cost of subchannel $k$ is

$$g^k(x_k) = \mathbb{E} \left[ \alpha_k f^\text{MSE} (\gamma_k) \right]$$

$$= \alpha_k g^\text{MSE}_{n_R-n_L+k} (x_k)$$

$$= \frac{\alpha_k}{(n_R-n_L+k-1)x_k}.$$

Let \( \bar{\alpha}_k = \log(\alpha_k/(n_R - n_L + k - 1)) \). Then by raising the cost function to the power of $q$ and using Design Formulation 3.2, we get

Design Formulation 3.6.

$$\minimize_{\bar{x}, \bar{p}} \sum_{k=1}^{n_L} e^{-q(\bar{x}_k-\bar{\alpha}_k)}$$

subject to $\bar{x} \preceq \bar{p} + \bar{\lambda}_T^l$,

$$\sum_{k=1}^{n_L} e^{\bar{p}_k} \leq P_{\text{max}}.$$

As already mentioned, the optimal ordering of $\bar{x}$ can be figured out in advance. Assume first that $\bar{\alpha}$ is decreasing, and that $(\bar{x}^*, \bar{p}^*)$ is a minimizer of Design Formulation 3.6. Then $(\bar{x}^*_l, \bar{p}^*_l)$ is feasible since the majorization constraint is insensitive to rearrangements on either side. This point is also optimal, since $\bar{x}^*_l - \bar{\alpha} \preceq \bar{x}^* - \bar{\alpha}$ (see [MOA11, Prop. 6.A.2]) and the objective $\sum_k e^{-q(\bar{x}_k-\bar{\alpha}_k)}$ is Schur-convex in $\bar{x} - \bar{\alpha}$ (since it is both symmetric and convex), and thereby order-preserving in “\( \preceq \)”.

When $\bar{\alpha}$ is a decreasing vector we may solve the problem using Design Formulation 3.5. Any other order of $\bar{\alpha}$ can be dealt with similarly. With the same reasoning as before, the conclusion is that it is optimal to let $\bar{x}$ be similarly ordered as $\bar{\alpha}$. Applying Design Formulation 3.5 with $\bar{\alpha}_l$ instead of $\bar{\alpha}$ gives the optimal $\bar{p}$ and $\bar{x}_l$, and the latter needs to be rearranged into the order of $\bar{\alpha}$ to be optimal for the original problem.

Interestingly, an almost closed-form solution exists to the specific optimization problem in Design Formulation 3.6. It can be computed using a simple convex-hull algorithm in $\mathcal{O}(n_L)$ time for a decreasing $\bar{\alpha}$. Establishing this result requires additional theory on optimization problems with majorization constraints, which we develop in Chapter 5. The solution to Design Formulation 3.6 is presented in Section 5.3.1.

### 3.3.3 Rate-Maximizing Design

Up to this point, we have considered optimal precoding for a fixed and predetermined rate allocation for the $n_L$ substreams. Next we consider
maximizing the achievable data rate that can be reliably maintained in fast fading, which is represented by the cost function

\[ F_{\gamma}(x) = \sum_{k=1}^{n_L} g^k(x_k), \]

where the average cost on the \( k \)th subchannel is given by

\[ g^k(x) = \mathbb{E}[-\log(1 + x\eta_n)], \quad (3.31) \]

where \( \eta_n \sim \chi^2_{2n}/2 \), and the explicit expression is given in (3.25). It is easily verified that \( g^k(x) \) is convex in \( x \), since (3.31) describes a convex mixture of convex functions in \( x \). However, in the logarithmic variable \( \bar{x} \), \( g^k(e^{\bar{x}}) \) is in fact concave by a similar argument. Hence, \( F_{\gamma}(e^{\bar{x}}) = \sum_{k=1}^{n_L} g^k(e^{\bar{x}_k}) \) is a Schur-concave composition of concave average costs. Properties of such cost functions were studied in the second part of Theorem 3.2, which contributes in the following way: For any feasible pair \((\bar{x}, \bar{p})\) in Design Formulation 3.2,

\[ F_{\gamma}(e^{\bar{x}}) \geq F_{\gamma}(e^{\bar{x}^\tau}) \geq F_{\gamma}(e^{(\bar{p} + \bar{\lambda}^T)^\tau}). \]

Since \( \bar{x}^* = (\bar{p} + \bar{\lambda}^T)^\tau \) is feasible and optimal given \( \bar{p} \), we may exchange the majorization constraint in Design Formulation 3.2 by an equality, and obtain in the following simplified optimization problem in original variables:

**Design Formulation 3.7.**

\[
\begin{align*}
\text{minimize} & \quad \sum_{k=1}^{n_L} g^k(x_k) \\
\text{subject to} & \quad x_{n_L-k+1} = p_k \lambda_{T,k}, \quad k = 1, \ldots, n_L, \\
& \quad \sum_{k=1}^{n_L} p_k = P_{\text{max}}.
\end{align*}
\]

We may simplify the problem further as \( \min_{p \geq 0} \sum_{k=1}^{n_L} g^{n_L-k+1}(p_k \lambda_{T,k}) \) subject to \( \sum_{k=1}^{n_L} p_k = P_{\text{max}} \).\textsuperscript{13} Such separable programs with a differentiable and strictly convex objective, which is the case here, can be very efficiently solved [Zip80].\textsuperscript{14}

The fact that the optimal \( x \) is increasing has an interesting interpretation. For equal-rate designs, as well as for MSE-based designs, the optimal

\textsuperscript{13}We then keep in mind that the minimizer \( p^* \) generates a minimizing pair \((x^*, p^*)\) in Design Formulation 3.7 by setting \( x^* = (p^* \odot \lambda^T)^\tau \). Note also that the number \( n_L^* \leq n_L \) of non-zero elements of \( p^* \) is the number of data streams to use in practice.

\textsuperscript{14}A familiar example in this class of optimization problems is the maximization of \( \sum_{k=1}^{n_L} \log(1 + p_k \lambda_k) \), which leads to the water-filling solution.
strategy is to equalize performance on the subchannels. From the perspective of maximizing the achievable data rate, on the other hand, this strategy turns out to be the worst possible. The optimal strategy is instead to allocate more power to subchannels decoded later, which already benefit more from interference cancellation.

3.4 Numerical Results

This section provides a few results and insights from simulations on DF systems with optimal precoders. We consider uncoded transmission over a $4 \times 4$ MIMO channel. In all cases, data is conveyed at a rate of 12 bits per transmission using complex symbols drawn from square QAM constellations. The performance is quantified by the joint error probability $Pr(\hat{s} \neq s)$ in (3.12) at each integer dB of the system SNR ($P_{\text{max}}$) between 15 dB and 35 dB.

Channel correlation is introduced using the exponential model [Loy01] through a single real parameter $\rho \in [0, 1]$ as

$$R_\rho = \begin{bmatrix}
1 & \rho & \rho^2 & \rho^3 \\
\rho & 1 & \rho & \rho^2 \\
\rho^2 & \rho & 1 & \rho \\
\rho^3 & \rho^2 & \rho & 1
\end{bmatrix}.$$  \hspace{1cm} (3.32)

The vector of eigenvalues $\lambda(R_\rho)$ is a function of $\rho$. In the simulations we use values of $\rho \in \{0, 0.3, 0.5, 0.7, 0.8, 0.9\}$, and increasing the value of $\rho$ should be interpreted as increasing the “amount” of channel correlation. Indeed, the choice $\rho = 0$ corresponds to an uncorrelated channel with $R_T = I$, and $\rho = 1$ amounts to the fully correlated case with $R_T$ being rank-1. As shown in [SW08, App. I], $\rho_0 \leq \rho_1$ implies $\lambda(R_{\rho_0}) \preceq \lambda(R_{\rho_1})$. Hence, we can relate the amount of transmit channel correlation present by the spread in the eigenvalues of $R_T$ as defined by the majorization relation. This quantification of transmit correlation has been used before in, e.g., [JB04].

3.4.1 Transmit Channel Correlation

We shall examine the impact of transmit channel correlation on performance. Here, we use a fixed, uniform rate allocation with three ($n_L = 3$) 16-QAM substreams. Figure 3.2 displays the performance for two designs in terms of the JEP. The only difference is that the “Prec. DF” design applies a JEP-optimized precoder, whereas “DF” refers to a simple diagonal equal-power precoder, which does not require any channel information at the transmitter. The power allocation of the optimized precoder is given
3.4. NUMERICAL RESULTS

Figure 3.2: Performance with and without precoding for different amounts of transmit correlation.

by solving Design Formulation 3.5, using (3.12) and (3.14) to evaluate the JEP.

Figure 3.2 considers three cases of different transmit correlation. By inspection, it is evident that without a precoder, the performance is degraded monotonically with increasing channel correlation. Interestingly, this is not the case if an optimal statistical precoder is used. In this case, moderate transmit correlation is actually beneficial. However, as the amount of correlation is increased further, performance is affected negatively in this case too. By comparing the performance gain of using a precoder at different degrees of correlation, we conclude that in this example system, the benefit of using a statistical precoder increases with the amount of transmit correlation present.

3.4.2 Rate Allocation

We now turn to studying the impact of substream rate allocation on performance. Again, we consider the uncoded setting with a fixed total rate of 12 bits per transmission. Bits are mapped into square QAM constellations using Gray encoding. This requires that an even number of bits is allocated to each subchannel. We have included moderately high transmit channel correlation with $\rho = 0.7$. 


Figure 3.3 presents the performance of five JEP-optimized designs with different rate allocations. The effect of rate allocation on performance is highly noticeable. At a JEP of 0.01, there is a 6 dB difference between the best and the worst examples of rate allocations. Also, it is evident that increasing rate allocations, which allocate more bits to subchannels detected later, are better than decreasing rate allocations, and the effect of having the rate allocation in the best order is not negligible. By performing an extensive search over the possible bit loads, it can be concluded that Figure 3.3 displays the optimal bit loads (either [6 6] or [2 4 6]) for SNRs between 15 dB and 30 dB.

### 3.4.3 MSE-based Designs

The MSE-based designs derived in Section 3.3.2 are in general not optimal when measuring performance in terms of the JEP, and this section studies the performance loss. Three different scenarios are considered. These differ in the choice of subchannel rate allocation and in the amount of transmit channel correlation present. An MSE-based design is optimized with respect to a \( q \)-norm of the expected MSEs according to (3.9). We consider the limiting cases with \( q = 1 \) and \( q = \infty \), corresponding to cost functionals

\[
\mathcal{F}_\gamma = \sum_{k=1}^{n_L} \mathbb{E} \left[ \alpha_k f^{\text{MSE}} (\gamma_k) \right] \quad \text{and} \quad \mathcal{F}_\gamma = \max \left\{ \mathbb{E} \left[ \alpha_k f^{\text{MSE}} (\gamma_k) \right] \right\}_{k=1}^{n_L},
\]

respectively. On each subchannel, the weight \( \alpha_k \) is related to the minimum
3.4. NUMERICAL RESULTS

distance of the constellation used, as described in Section 3.2.2. Our two weighted-MSE based designs, termed “wsum-MSE” and “wmax-MSE” in Figure 3.4, are compared with the JEP-optimized precoders.

For the cases studied in Figure 3.4, we conclude that in general our MSE-based designs perform well even in terms of the JEP. With rate allocation [6 4 2] and an uncorrelated channel ($\rho = 0$), the “wsum-MSE” is close to optimal in the low-SNR region, while “wmax-MSE” is close to optimal for high SNRs. With rate allocation [2 4 6] and minor transmit channel correlation ($\rho = 0.3$), “wmax-MSE” performs best for low SNRs instead, and the “wsum-MSE” precoder is close to optimal over the entire SNR range studied. When transmit correlation is increased to $\rho = 0.8$, and a uniform rate allocation is used, the two MSE-based designs display identical performance, with a common performance loss for high SNRs.

3.4.4 Detection Order

It is widely known that the choice of detection order for a DF receiver can have an important effect on system performance, which is elaborated on in, e.g., [WFGV98] and [PV04]. This section explores numerically the effect of optimal ordering in the previously considered case of a $4 \times 4$ system with three 16-QAM substreams. A transmit channel correlation of $\rho = 0.7$ is included.
We compare the performance of four designs with a DF receiver: with and without a JEP-optimized precoder, and in either case with and without an optimal detection order. The optimal detection order is found by an extensive search over all detection orders, with the instantaneous JEP, $1 - \prod_k \left[ 1 - f_{\text{SEP}}(\gamma_k) \right]$, as objective [PV04]. The resulting computational complexity at the receiver is in general prohibitive, and it is common to use more efficient, but suboptimal, ordering schemes. In this numerical example, with $n_L = 3$, the computational complexity is manageable, and we therefore use the optimal ordering.

The JEP-optimal precoder for the DF receiver with optimal ordering is not known. Therefore, the two precoded transceivers considered here both use the precoder optimized for the ordinary DF receiver. Note that this need not be the optimal choice for the optimally ordering receiver.

The SNR characterization in Theorem 3.1 is not valid for an optimally ordering DF receiver. In order to evaluate the performance we resort to a Monte Carlo simulation. The results in Figure 3.5 are based on a total number of $10^6$ channel realizations. For each channel realization and transceiver, the instantaneous JEP is determined as in [PV04], and the JEP is computed by taking the average over all channel realizations.

Figure 3.5 displays the performance of the four schemes. Without precoding, optimal detection ordering has a highly noticeable positive impact.
on performance. However, in this example there is a larger gain in using an optimal precoder instead. Combining precoding and optimal ordering results, naturally, in the best performance. Note that with this precoder, the additional gain of using an ordered receiver is marginal. This is not surprising, since the statistical precoder optimized for a fixed detection order should render the same detection order optimal for most channel realizations. In this way, the gain associated with optimal ordering is reduced. This gain might, however, be restored if the DF receiver with optimal ordering is equipped with a statistical precoder tailored to this receiver design.

3.5 Summary

We have investigated the problem of designing a statistical precoder to assist the ZF-DF equalizer in a point-to-point MIMO communication system. We posed the problem of maximizing system performance subject to a transmit-power constraint, and presented a novel, efficiently solvable reformulation of the optimization problem in Design Formulation 3.4 with applicability to all marginal convex cost functions (Def. 3.2). Two special instances of the precoder-optimization problem admitted a further simplified problem in Design Formulation 3.5: systems with a uniform subchannel rate allocation, as well as the case when measuring performance using an MSE-based cost function. The behavior of DF transceivers was evaluated numerically, characterizing optimized precoders in several respects. Examples were provided revealing an increasing benefit of precoder optimization with the amount of transmit channel correlation present, as well as the additional importance of combining precoder optimization with proper subchannel rate allocation. In our study, precoder optimization for a particular detection order left little room for further enhancement using the more complex DF receiver with optimal ordering. Finally, MSE-based precoders displayed only a marginal performance degradation to the optimal precoders. This conclusion serves to motivate an in-depth study of the MSE-based precoder optimization problem, which we provide in Chapter 5, showing that such problems can be efficiently solved without resorting to generic algorithms of convex optimization.
3. A Collection of Proofs

3. A.1 Proof of Lemma 3.1

The signal model in (3.2) is particularly simple under the ZF constraint:
\[ \bar{s} = s + W^H n. \]

The MSEs to be minimized are the diagonal entries of the MSE matrix
\[ \mathbb{E} \left[ (\bar{s} - s)(\bar{s} - s)^H \right] = W^H W, \]
and the vector of diagonal entries is denoted by \( d(W^H W) \).

We first find the MSE-minimizing feedforward matrix \( W \) for a given feedback matrix \( B \). Using \( QL = HP \) and the ZF constraint \( W^H QL = B + I \), the MSE matrix is easily expanded as
\[
W^H W = [W^H - (B + I)L^{-1}Q^H] [W^H - (B + I)L^{-1}Q^H]^H \\
+ (B + I)(L^H L)^{-1}(B + I)^H. \tag{3.33}
\]

The first term on the right-hand side is a positive semi-definite matrix. Its contribution to the MSE vector \( d(W^H W) \) is hence a vector with non-negative components, and it follows that
\[
d(W^H W) \geq d\left((B + I)(L^H L)^{-1}(B + I)^H\right),
\]
where the inequality applies componentwise. Equality is attained by setting
\[
W^H = (B + I)L^{-1}Q^H \tag{3.34}
\]
in (3.33), and we note that the ZF constraint \( W^H QL = B + I \) is satisfied.

The optimal feedback matrix can be similarly obtained. With some manipulations,
\[
(B + I)(L^H L)^{-1}(B + I)^H = (B + I - D_L^{-1}L)(L^H L)^{-1}(B + I - D_L^{-1}L)^H \\
+ D_L^{-1}L^{-H}(B + I)^H + (B + I)L^{-1}D_L^{-1} \\
- D_L^{-2}
\]
The first term is positive semi-definite, and the diagonal entries of the second and third terms do not depend on \( B \) if it is strictly lower triangular, as required by the detection order. In fact, both of these terms have diagonal entries that equal \( d(D_L^{-2}) \). It follows that
\[
d\left((B + I)(L^H L)^{-1}(B + I)^H\right) \geq d\left(D_L^{-2}\right). \tag{3.35}
\]
Equality is attained by setting $B = B_{ZF} = D_L^{-1}L - I$, which indeed is a strictly lower triangular matrix. With this feedback matrix in (3.34) we also obtain the desired expression $W_{ZF} = QL^{-H}(B_{ZF} + I)^H = QD_L^{-1}$. The proof is concluded by noting that all MSEs have been minimized independently of the others.

3.A.2 Proof of Theorem 3.1

To prove the theorem we need the following factorization rule for Cholesky elements, which will be used in subsequent chapters as well.

**Lemma 3.4.** For a positive semidefinite matrix $T \in \mathbb{C}^{M \times M}$ and a tall matrix $S \in \mathbb{C}^{M \times N}$ having a QL decomposition $QL = S$,

$$l^2(S^HTS) = l^2(Q^HTQ) \odot l^2(S^HS). \tag{3.36}$$

**Proof.** Using $QL = S$ and the Cholesky decomposition

$$L_1^HL_1 = Q^HTQ,$$

it follows that

$$S^HTS = L^H_1L^H_1L_1L.$$ Defining $L_2 = L_1L$, we find that $L_2$ is lower triangular and $S^HTS = L_2^HL_2$ is a Cholesky decomposition. The diagonal of $L_2$ is the componentwise product of the diagonals of $L_1$ and $L$. This fact translates into (3.36) in terms of Cholesky elements as defined in Definition 3.1. □

With $H \sim ZR_T^{1/2}$ and $\gamma = l^2(P^HH^HHP)$ in (3.4) and (3.7), respectively, it follows from Lemma 3.4 that

$$\gamma \sim l^2(Q^HZ^HZQ) \odot l^2(P^HR_T^HP),$$

where the semi-unitary matrix $Q \in \mathbb{C}^{n_T \times n_L}$ stems from the QL decomposition $QL = R_T^{1/2}P$. It follows from the unitary invariance of $Z$ that, irrespective of $Q$, $ZQ \sim \tilde{Z}$, where $\tilde{Z} \in \mathbb{C}^{n_R \times n_L}$ has i.i.d. $CN(0,1)$ entries. The joint distribution of $l^2(\tilde{Z}^HZ)$ is well-known (see, e.g., [Goo63] and [TV04, Lemma 2.1]): the components are independent and $l^2_k(\tilde{Z}^HZ) \sim \chi^2_{2(n_R - n_L + k)/2}$, and the result follows.
3.A.3 Proof of Theorem 3.2

The proof of Theorem 3.2 starts with a lemma. In order to keep the notation neat, we define a class of average costs as

$$\bar{g}_i(\bar{x}) = \mathbb{E} \left[ f \left( e^{\bar{x}} \eta_i \right) \right], \quad i \geq 1,$$

(3.37)

where $\eta_i = \chi_{2i}^2 / 2$ follows a Gamma distribution with PDF

$$p_i(\eta) = \frac{1}{\Gamma(i)} \eta^{i-1} e^{-\eta}, \quad \eta \geq 0.$$

(3.38)

Under the SNR distribution in Theorem 3.1, and with a common inner cost function $f = f^1 = \ldots = f^{n_L}$ on the subchannels, these relate to the average subchannel costs as $g_k(\bar{x}) = \bar{g}_{n_R-n_L+k}(\bar{x})$, and

$$\mathcal{F}_\gamma(e^{\bar{x}}) = \mathcal{G} (\{ \bar{g}_{n_R-n_L+k}(\bar{x}) \}_{k=1}^{n_L}).$$

(3.39)

Lemma 3.5. Assume that $f: \mathbb{R}_+ \to \mathbb{R}$ is chosen such that, for some $i_0 \geq 1$,

$$\mathbb{E} \left[ |f(e^{\bar{x}} \eta)| \right] < \infty$$

(3.40)

for all $\bar{x} \in \mathbb{R}$ and $i \geq i_0$. Then each $\bar{g}_i(\bar{x}) = \mathbb{E} [ f( e^{\bar{x}} \eta_i ) ]$ is infinitely differentiable on $\mathbb{R}$, with

$$\bar{g}_i'(\bar{x}) = i \left[ \bar{g}_{i+1}(\bar{x}) - \bar{g}_i(\bar{x}) \right].$$

(3.41)

Proof. First note that is suffices to prove (3.41) for $i \geq i_0$, since infinite differentiability follows by induction. Fix $i \geq i_0$ and $\bar{x} \in \mathbb{R}$. Then

$$\bar{g}_i(\bar{x}) = \int_0^\infty f \left( e^{\bar{x}} \eta \right) p_i(\eta) d\eta = \int_0^\infty f (t) e^{-\bar{x}} p_i(e^{-\bar{x}} t) dt.$$

Using only (3.38), it is straightforward to show that

$$\frac{\partial}{\partial \bar{x}} \phi_i(\bar{x}, t) = i \left[ \phi_{i+1}(\bar{x}, t) - \phi_i(\bar{x}, t) \right],$$

(3.42)

from which we conclude that

$$\int_0^\infty \frac{\partial}{\partial \bar{x}} \phi_i(\bar{x}, t) dt = i \left[ \bar{g}_{i+1}(\bar{x}) - \bar{g}_i(\bar{x}) \right].$$

It remains to verify that $\bar{g}_i'(\bar{x})$ exists, and that

$$\bar{g}_i'(\bar{x}) = \frac{d}{d\bar{x}} \int_0^\infty \phi_i(\bar{x}, t) dt = \int_0^\infty \frac{\partial}{\partial \bar{x}} \phi_i(\bar{x}, t) dt.$$

(3.43)
For the sake of brevity, we only outline a formal verification of this claim. By [Ada99, p. 802], (3.43) follows if there exist \(a, b \in \mathbb{R}\) such that \(a < a \) and a function \(\psi_i(t)\) such that \(\left| \frac{\partial^2}{\partial x^2} \phi_i(\bar{x}, t) \right| \leq \psi_i(t)\) on \(a < a \), and such that \(\int_0^\infty \psi_i(t) \, dt < \infty\). For any \(a < a < b\), we can express \(\left| \frac{\partial^2}{\partial x^2} \phi_i(\bar{x}, t) \right|\) by repeated use of (3.42), and take \(\psi_i(t)\) as its supremum over \(a < a < b\). The integral over \(\psi_i(t)\) can then be bounded using moments on the form in (3.40) for \(i \geq i_0\).

The remainder of the proof uses two alternate forms of majorization: weak submajorization \(\prec_w\) and weak supermajorization \(\prec_w\) defined as follows. For \(a, b \in \mathbb{R}^N\), \(a \prec_w b\) if and only if \(\sum_{k=1}^{n} a_{jk} \leq \sum_{k=1}^{n} b_{jk}\) for all \(n = 1, \ldots, N\). Similarly, \(a \prec_w b\) if and only if \(\sum_{k=1}^{n} a_{jk} \geq \sum_{k=1}^{n} b_{jk}\) for all \(n = 1, \ldots, N\).

The first part of Theorem 3.2 assumes an increasing, Schur-convex outer cost function \(\mathcal{G}\) and convex functions \(g_k(e^\bar{x}) = \tilde{g}_{nR-nL+k}(\bar{x})\) for \(k = 1, \ldots, nL - 1\). [MOA11, Thm. 3.A.8] concludes that any increasing, Schur-convex function is order-preserving in the weak majorization \(\prec_w\). By (3.39), the desired result in (3.28) then follows if

\[
\{\tilde{g}_{nR-nL+k}(\bar{x}_{1:k})\}_{k=1}^{nL} \prec_w \{\tilde{g}_{nR-nL+k}(\bar{x}_{1:k})\}_{k=1}^{nL}.
\]

(3.44)

Consider first the case \(n_L = 2\). We note that if \(f\) is decreasing, then each \(\tilde{g}_i(\bar{x})\) in (3.37) is decreasing, and (3.41) further implies that \(\tilde{g}_i(\bar{x}) \geq \tilde{g}_{i+1}(\bar{x})\) for any \(\bar{x}\). If \(\bar{x} \geq \bar{x}_2\), then

\[
\max \{\tilde{g}_i(\bar{x}_2), \tilde{g}_{i+1}(\bar{x}_2)\} = \tilde{g}_i(\bar{x}_2) \geq \max \{\tilde{g}_i(\bar{x}_1), \tilde{g}_{i+1}(\bar{x}_2)\}.
\]

(3.45)

It also follows from (3.41) that

\[
\tilde{g}_i(\bar{x}_2) + \tilde{g}_{i+1}(\bar{x}_1) - [\tilde{g}_i(\bar{x}_1) + \tilde{g}_{i+1}(\bar{x}_2)] = \frac{1}{i} \left[ \tilde{g}_i(\bar{x}_1) - \tilde{g}_i(\bar{x}_2) \right].
\]

(3.46)

Since \(\tilde{g}_i(\bar{x})\) is convex, \(\tilde{g}_i(\bar{x}_1) \geq \tilde{g}_i(\bar{x}_2)\). A non-negative left-hand side of (3.46) in conjunction with (3.45) establishes the weak submajorization

\[
\{\tilde{g}_i(\bar{x}_1), \tilde{g}_{i+1}(\bar{x}_2)\} \prec_w \{\tilde{g}_i(\bar{x}_2), \tilde{g}_{i+1}(\bar{x}_1)\}
\]

(3.47)

for \(i \geq i_0\) in Lemma 3.5. Setting \(i = nR - nL + 1\) results in (3.44) for \(n_L = 2\).

In order to prove (3.44) for \(n_L \geq 3\), we can make repeated use of (3.47). First note that for any \(a_1, \ldots, a_J\) and \(j\) such that \(0 \leq j \leq J\), \((y_1, y_2) \prec_w (z_1, z_2)\) implies that [MOA11, Prop. 5.A.7]

\[
(a_1, \ldots, a_j, y_1, y_2, a_{j+1}, \ldots, a_J) \prec_w (a_1, \ldots, a_j, z_1, z_2, a_{j+1}, \ldots, a_J).
\]

(3.48)
Let \( \{\tilde{x}^0, \ldots, \tilde{x}^N\} \) denote a sorting process of \( \tilde{x} = \tilde{x}^0 \) into \( \tilde{x}_1 = \tilde{x}^N \) such that for each \( n \) there is an \( m \) such that
\[
\tilde{x}^{n+1} = (\tilde{x}^n_1, \ldots, \tilde{x}^n_{m-1}, \tilde{x}^n_m, \tilde{x}^n_{m+1}, \ldots, \tilde{x}^n_{n_L}),
\]
where \( \tilde{x}^n_{m+1} > \tilde{x}^n_m \). It follows by (3.47) that
\[
\{\tilde{g}_{nR-nL+m}(\tilde{x}^n_m), \tilde{g}_{nR-nL+m+1}(\tilde{x}^n_m)\} \prec \{\tilde{g}_{nR-nL+m}(\tilde{x}^n_m), \tilde{g}_{nR-nL+m+1}(\tilde{x}^n_m)\},
\]
and furthermore by (3.48) that
\[
\{\tilde{g}_{nR-nL+k}(\tilde{x}^{n+1}_k)\}_{k=1}^{nL} \prec \{\tilde{g}_{nR-nL+k}(\tilde{x}^n_k)\}_{k=1}^{nL}.
\]
Using the transitivity of \( \prec \), we end up with (3.44), which concludes the first part of the proof.

The second part of Theorem 3.2 assumes a concave outer cost function \( G \) and concave functions \( g^k(e^x) = \tilde{g}_{nR-nL+k}(\tilde{x}) \) for \( k = 1, \ldots, n_L - 1 \). For \( \tilde{x}_1 \geq \tilde{x}_2 \),
\[
\min \{\tilde{g}_i(\tilde{x}_2), \tilde{g}_{i+1}(\tilde{x}_1)\} = \tilde{g}_{i+1}(\tilde{x}_1) \leq \min \{\tilde{g}_i(\tilde{x}_1), \tilde{g}_{i+1}(\tilde{x}_2)\}. \tag{3.49}
\]
Combining (3.49) with (3.46) for a concave \( \tilde{g}_i(\tilde{x}) \) then results in
\[
\{\tilde{g}_i(\tilde{x}_1), \tilde{g}_{i+1}(\tilde{x}_2)\} \prec \{\tilde{g}_i(\tilde{x}_2), \tilde{g}_{i+1}(\tilde{x}_1)\}.
\]
With trivial modifications of the above procedure we can similarly prove that
\[
\{\tilde{g}_{nR-nL+k}(\tilde{x}^1_k)\}_{k=1}^{nL} \prec \{\tilde{g}_{nR-nL+k}(\tilde{x}^2_k)\}_{k=1}^{nL} \tag{3.50}
\]
for \( n_L \geq 2 \). In this case \( G \) is Schur-concave and increasing, \( y \prec w z \) implies \( G(z) \leq G(y) \), and the desired result (3.29) follows. In order to prove that \( F^R(e^x) \) becomes Schur-concave on the set of increasing vectors \( \tilde{x} \), it suffices to verify that the result [MOA11, 5.A.3] applies: In this context it states that for two increasing vectors \( \tilde{x}^1_i \) and \( \tilde{x}^2_i \) such that \( \tilde{x}^1_i \preceq \tilde{x}^2_i \),
\[
\{\tilde{g}_{nR-nL+k}(\tilde{x}^1_k)\}_{k=1}^{nL} \prec \{\tilde{g}_{nR-nL+k}(\tilde{x}^2_k)\}_{k=1}^{nL}
\]
holds if the following two conditions are satisfied for \( n_R - n_L < i < n_R \):
\[
\tilde{g}_i(\tilde{x}) \geq \tilde{g}_{i+1}(\tilde{x}), \quad \forall \tilde{x} \in \mathbb{R}, \tag{3.51}
\]
\[
0 \geq \tilde{g}_i'(a) \geq \tilde{g}_{i+1}'(b), \quad a \leq b. \tag{3.52}
\]
We have already concluded that \( \tilde{g}_i(\tilde{x}) \) is decreasing in both \( \tilde{x} \) and \( i \). Using (3.41) and the fact that each \( \tilde{g}_i(\tilde{x}) \) is concave it follows that
\[
\tilde{g}_i'(a) = \tilde{g}_{i+1}'(a) - \frac{1}{i} \tilde{g}_i''(a) \geq \tilde{g}_{i+1}'(a) \geq \tilde{g}_{i+1}'(b).
\]
This concludes the proof.
3.A.4 Proof of Theorem 3.3

We shall use the SVD $\mathbf{P} = \mathbf{U} \mathbf{D} (\sqrt{\mathbf{P}}) \mathbf{V}^H$ throughout the proof, which starts with a lemma.

**Lemma 3.6.** Given two $n_L$-vectors $\mathbf{x}$ and $\mathbf{z}$ with real, positive components, there is a unitary matrix $\mathbf{V} \in \mathbb{C}^{n_L \times n_L}$ such that $\mathbf{x} = \mathbf{I}^2 (\mathbf{V} \mathbf{D} (\mathbf{z}) \mathbf{V}^H)$ if and only if $\log(\mathbf{x}) \preceq \log(\mathbf{z})$.

**Proof.** This is a consequence of the generalized triangular decomposition (GTD) theorem [JHL08], which also provides an algorithm to determine $\mathbf{V}$ given $\mathbf{x}$ and $\mathbf{z}$. \[\square\]

It follows by Lemma 3.6 that

$$\log(\mathbf{I}^2 (\mathbf{P}^H \mathbf{R}_T \mathbf{P})) \preceq \log(\mathbf{J}) (\mathbf{P}^H \mathbf{R}_T \mathbf{P}),$$

where $\mathbf{J}$ denotes a decreasing vector of eigenvalues. Next, [MOA11, Thm. 9.H.1.a] further implies that

$$\log(\mathbf{J}) (\mathbf{P}^H \mathbf{R}_T \mathbf{P}) \preceq \log(\mathbf{P}_+ \mathbf{J}) + \log(\mathbf{J}) (\mathbf{U}^H \mathbf{R}_T \mathbf{U}),$$

where $\mathbf{P}_+ = \mathbf{J} (\mathbf{P}^H \mathbf{P})$. Moreover, by [MOA11, 9.H.3]

$$\log(\mathbf{J}) (\mathbf{U}^H \mathbf{R}_T \mathbf{U}) \prec_w \log(\mathbf{J}_T^L),$$

where $\mathbf{J}_T^L \in \mathbb{R}^{n_L}$ denotes the decreasing vector of the $n_L$ largest eigenvalues of $\mathbf{R}_T$, and weak submajorization “$\prec_w$” was defined in Appendix 3.A.3. Combining the orderings for $\mathbf{x} = \mathbf{I}^2 (\mathbf{P}^H \mathbf{R}_T \mathbf{P})$ gives

$$\log(\mathbf{x}) \preceq_w \log(\mathbf{P}_+) + \log(\mathbf{J}_T^L). \quad (3.54)$$

The reverse implication need not be true; given $\mathbf{x}$ and $\mathbf{p}$ satisfying (3.54), it may not be possible to find a precoder $\mathbf{P}$ satisfying $(\mathbf{x}, \mathbf{p}_+) = (\mathbf{I}^2 (\mathbf{P}^H \mathbf{R}_T \mathbf{P}), \mathbf{J} (\mathbf{P}^H \mathbf{P}))$. We shall ignore this fact for the moment, fix $\mathbf{p}$ and consider the set of vectors $\mathbf{x}$ satisfying (3.54). For any $\mathbf{x}^\ast$ in this set, there is a $\mathbf{y}$ such that $\mathbf{x}^\ast \leq \mathbf{y}$ componentwise, and (see [MOA11, 5.A.9])

$$\log(\mathbf{y}) \preceq \log(\mathbf{P}_+) + \log(\mathbf{J}_T^L).$$

Since the cost function $\mathcal{F}_\gamma (\mathbf{x})$ of Design Formulation 3.1 is decreasing in each argument by Lemma 3.2, we must have $\mathcal{F}_\gamma (\mathbf{y}) \leq \mathcal{F}_\gamma (\mathbf{x}^\ast)$. In conclusion, any $\mathbf{x}$ that is not on the “boundary” $(\preceq)$ of (3.54) is not relevant for optimization, provided that there is a $\mathbf{P}$ associated with any $\mathbf{x}$ on the boundary. This entire set can be realized using the proposed $\mathbf{U}$ in Theorem 3.3 and a decreasing $\mathbf{p}$. Indeed, this choice renders $\mathbf{P}^H \mathbf{R}_T \mathbf{P} = \mathbf{V} \mathbf{D} \left( \mathbf{P} \odot \mathbf{J}_T^L \right) \mathbf{V}^H$, and the proof is concluded by Lemma 3.6.
Chapter 4

Performance Measures in Correlated MIMO Channels

The previous chapter developed an optimization framework for statistical precoding for the ZF-DF transceiver in fading MIMO channels, focusing on the transmit-correlated Rayleigh channel model. This chapter extends the range of the framework to include the separable-correlation Rayleigh model with both transmit and receive correlation, as well as the double-scattering model. The main issue, however, is that analytic expressions for common performance measures are not available in the literature. This chapter presents means to obtain explicit expressions for a variety of such performance measures for the two channel models.

4.1 SNR Characterization

The starting point in this chapter is the characterization of subchannel SNRs for the ZF-DF transceiver with statistical precoding, as presented in Theorem 3.1. For the transmit-correlated Rayleigh-fading model, it was shown that the subchannel SNR vector $\gamma \in \mathbb{R}^{n_L}$ can be factored as

$$
\gamma \sim \bar{\gamma} \odot x,
$$

in terms of a random component, the normalized SNR vector $\bar{\gamma}$, and a deterministic component, the power allocation vector $x = l^2 (P^H R T P)$. The components of the normalized SNR vector $\bar{\gamma}$ were independent, scaled chi-squared distributed random variables with different degrees of freedom.

---

1These models are described in Section 2.1.1.

2Recall Definition 3.1, stating that for a positive semidefinite matrix $T \in \mathbb{C}^{N \times N}$, the vector of Cholesky elements $l^2 (T) \in \mathbb{R}^N_+$ is comprised of the squared diagonal elements of the lower-triangular matrix $L$ in the Cholesky decomposition $L^H L = T$. 
CHAPTER 4. PERFORMANCE IN CORRELATED CHANNELS

This was a mark left by the statistical channel model adopted, and the normalized subchannel SNRs could not be altered by means of statistical precoding; the role of precoding was confined to the subchannel power allocation vector \( x = l^2 (P^H R_T P) \). This was a key feature of the framework for statistical precoding in the preceding chapter.

Interestingly, the important factorization in (4.1) holds for more generally correlated MIMO channel models. We particularize the treatment here to the following two channel models that were detailed in Section 2.1.1:

- **Separable-Correlation Rayleigh Model**: \( H = R_R^{1/2} Z R_T^{1/2} \),
- **Double-Scattering Model**: \( H = R_R^{1/2} Z_R R_{S}^{1/2} Z_S R_T^{1/2} \),

where the matrices \( Z \in \mathbb{C}^{n_R \times n_L} \), \( Z_R \in \mathbb{C}^{n_R \times n_S} \), and \( Z_S \in \mathbb{C}^{n_S \times n_T} \) all have i.i.d. \( \mathcal{CN}(0, 1) \) entries.

**Theorem 4.1.** Under the separable-correlation Rayleigh model and \( n_L \leq \min(n_R, n_T) \), the subchannel SNR vector can be factored as

\[
\gamma_{\text{Ray}} \sim \bar{\gamma}_{\text{Ray}} \circ x, \tag{4.2}
\]

in terms of a random normalized SNR vector \( \bar{\gamma}_{\text{Ray}} \), and a deterministic power allocation vector \( x = l^2 (P^H R_T P) \). The normalized SNR distribution is characterized by

\[
\bar{\gamma}_{\text{Ray}} \sim l^2 (\bar{Z}^H R_R \bar{Z}), \tag{4.3}
\]

where \( \bar{Z} \in \mathbb{C}^{n_R \times n_L} \) has i.i.d. \( \mathcal{CN}(0, 1) \) entries. Similarly, under the correlated double-scattering model and \( n_L \leq \min(n_R, n_S, n_T) \), the subchannel SNR vector can be factored as

\[
\gamma_{\text{DS}} \sim \bar{\gamma}_{\text{DS}} \circ x, \tag{4.4}
\]

where again \( x = l^2 (P^H R_T P) \), and the normalized SNR distribution is characterized by

\[
\bar{\gamma}_{\text{DS}} \sim l^2 (\bar{Z}_R^H R_R \bar{Z}_R) \circ l^2 (\bar{Z}_S^H R_S \bar{Z}_S), \tag{4.5}
\]

with \( \bar{Z}_R \in \mathbb{C}^{n_R \times n_L} \) and \( \bar{Z}_S \in \mathbb{C}^{n_S \times n_L} \) being statistically independent matrices, each with i.i.d. \( \mathcal{CN}(0, 1) \) entries.

**Proof.** The proof is located in Appendix 4.A.1. \( \square \)

The objective in this chapter is to obtain exact or approximate expressions for common performance measures under the two correlated MIMO
4.2. PERFORMANCE IN GENERALIZED GAMMA FADING

channel models. We will specifically consider average performance for each subchannel on the form

\begin{align}
g_n^{\text{Ray}}(x) &= \mathbb{E} \left[ f(\hat{\gamma}_n^{\text{Ray}} x) \right], \\
g_n^{\text{DS}}(x) &= \mathbb{E} \left[ f(\hat{\gamma}_n^{\text{DS}} x) \right],
\end{align}

for \( n = 1, \ldots, n_L \) with respect to an instantaneous performance measure \( f : \mathbb{R}_+ \to \mathbb{R} \). These are functions of the power \( x \in \mathbb{R}_+ \) allocated to the subchannel. With such expressions at hand, it is possible to evaluate the performance of using a certain precoder \( P \), using the explicit relation \( x = l^2(P^H R_T P) \) between the subchannel power allocation vector \( x \) and the precoder \( P \). This suffices to employ the optimization methods in the previous chapter when evaluating system performance based on the marginal distributions of the SNRs \( \gamma_1, \ldots, \gamma_{n_L} \) (see Definition 3.2).

The chapter is organized as follows. In Section 4.2 we focus on the subchannel last in line to be decoded (\( n = n_L \)), and present exact analytical expressions for several common performance measures under both Rayleigh and double-scattering fading. In Section 4.3, we consider the other subchannels and present PDF approximations for the normalized SNRs. These approximations are based on the same family of distributions as that of the last subchannel. Hence, we can use the exact expressions previously derived as approximate ones for the other subchannels as well. In Section 4.4 we finally evaluate numerically the proposed approximations and conclude that they are accurate, and hence suitable for precoder optimization.

### 4.2 Performance Under the Generalized Gamma Distribution

Consider the last subchannel in line to be decoded under the correlated Rayleigh channel model. The last component \( \hat{\gamma}_{n_L}^{\text{Ray}} \) of the vector \( \hat{\gamma}^{\text{Ray}} \sim l^2(Z^H R_R Z) \) is particularly easy to express. By the very definition of Cholesky elements, \( \hat{\gamma}_{n_L}^{\text{Ray}} \sim l_{n_L n_L}^2 \), where \( l_{n_L n_L} \) is the lower right entry of the lower-triangular matrix \( L \in \mathbb{C}^{n_L \times n_L} \) in the Cholesky decomposition \( L^H L = Z^H R_R Z \). It trivially follows that \( l_{n_L n_L}^2 \) is nothing but the lower right entry of the matrix \( Z^H R_R Z \). The normalized SNR \( \hat{\gamma}_{n_L}^{\text{Ray}} \) is then distributed as

\[ \hat{\gamma}_{n_L}^{\text{Ray}} \sim \mathcal{Z}^H R_R \mathcal{Z}, \]

with \( \mathcal{Z} \in \mathbb{C}^{n_R} \) having i.i.d. \( \mathcal{CN}(0,1) \) elements. Since \( \mathcal{Z} \sim Q \mathcal{Z} \) for any unitary matrix \( Q \), we may regard \( \hat{\gamma}_{n_L}^{\text{Ray}} \) as a weighted sum of the squared components of \( \mathcal{Z} \) as

\[ \hat{\gamma}_{n_L}^{\text{Ray}} \sim \lambda_{R,1} |z_1|^2 + \ldots + \lambda_{R,n_R} |z_{n_R}|^2, \]
with weights $\lambda_{R,1} \geq \ldots \geq \lambda_{R,n_R} > 0$ being the eigenvalues of $R_R$. It is well known that $|z_1|^2, \ldots, |z_{n_R}|^2$ are i.i.d. standard exponential random variables, and we shall refer to the family of distributions on the form (4.9) as generalized gamma distributions. These distributions play an important role for the double-scattering channel model as well. The normalized SNR on the last subchannel similarly becomes a product of generalized gamma distributions, since $\tilde{\gamma}_{DS}^{n_l}$ can be expressed as

$$\tilde{\gamma}_{DS}^{n_l} \sim z_R^H R_R z_R \cdot z_S^H R_S z_S,$$

where $z_R \in \mathbb{C}^{n_R}$ and $z_S \in \mathbb{C}^{n_S}$ are independent with i.i.d. $\mathcal{CN}(0,1)$ elements. Hence, as a natural first step towards expressions for average performance on the form (4.6) and (4.7), we analyze the generalized gamma distribution in some detail next.

### 4.2.1 The Generalized Gamma Distribution

**Definition 4.1 (Generalized Gamma Distribution).** Given $\lambda_1 \geq \ldots \geq \lambda_N > 0$ and i.i.d. standard exponential random variables $X_1, \ldots, X_N$ (with PDF $p(z) = e^{-z}, z \geq 0$), the weighted sum

$$\tilde{\gamma} = \lambda_1 X_1 + \ldots + \lambda_N X_N \sim \Gamma(\lambda_1, \ldots, \lambda_N)$$

is said to follow a generalized gamma distribution of order $N$ with parameters $\lambda_1, \ldots, \lambda_N$, and with the corresponding PDF denoted as $p_N^{\Gamma}(z; \lambda_1, \ldots, \lambda_N)$.

If the parameters are distinct ($\lambda_1 > \ldots > \lambda_N$), the density can be compactly expressed as [Cox62, p.17]

$$p_N^{\Gamma}(z; \lambda_1, \ldots, \lambda_N) = \sum_{n=1}^{N} \frac{\lambda_n^{N-2} e^{-z/\lambda_n}}{\prod_{k \neq n}(\lambda_n - \lambda_k)}. \quad (4.11)$$

With identical parameters $\lambda_1 = \ldots = \lambda_N = \lambda$, Definition 4.1 describes a classic gamma distribution with shape parameter $N$ and scale parameter $\lambda$, corresponding to the PDF

$$p_N^{\Gamma}(z; \lambda, \ldots, \lambda) = \frac{z^{N-1} e^{-z/\lambda}}{(N-1)! \lambda^N}.$$ 

In the general case, with repeated values for some of the parameters, the PDF is known but has a quite intricate structure. An early reference is

---

3As shown in the following, this is really a generalization of the gamma distribution for integer shape parameters. Note that there are also other common generalizations of the gamma distribution in the literature.
4.2. PERFORMANCE IN GENERALIZED GAMMA FADING

[Mat82] within mathematical statistics, but explicit expressions can also be found in [AM97, Coe98, KST06] originating from three different disciplines: reliability theory, multivariate analysis, and communications, respectively. Recently, an integral representation in terms of the Meijer G function was presented in [AYAK12]. Herein, we adopt a different approach to deal with the general case, which is based on a novel characterization of the PDF. For this purpose, we introduce the concept of divided differences as defined next.

Definition 4.2 (Divided Difference). Given \( \lambda_1 \geq \ldots \geq \lambda_N \) and a real-valued function \( h(\lambda) \), the divided difference of \( h \) at \( \lambda_1, \ldots, \lambda_N \) is denoted by \( \Delta(\lambda_1, \ldots, \lambda_N)h \) and is recursively defined as follows, where \( h^{(k)}(\lambda) \) denotes the \( k \)-th derivative of \( h(\lambda) \):

\[
\Delta(\lambda_i, \ldots, \lambda_j)h = \begin{cases} 
\Delta(\lambda_i, \ldots, \lambda_{j-1})h - \Delta(\lambda_{i+1}, \ldots, \lambda_j)h & \text{if } \lambda_i \neq \lambda_j, \\
\left(\frac{h^{(j-i)}(\lambda_i)}{\lambda_i - \lambda_j}\right) & \text{if } \lambda_i = \lambda_j,
\end{cases}
\]

(4.12)

for \( 1 \leq i \leq j \leq N \), provided that the necessary derivatives of \( h(\lambda) \) exist.

The concept of divided differences, dating back to Newton in 1687, is a versatile one; the definition provided here is only one in a plurality of descriptions. Other facets of the concept include the representation as a contour integral, a ratio of determinants, and as coefficients arising in polynomial interpolation. For a contemporary exposition on divided differences, we refer to [dB05].

By inspection of Definition 4.2 in the case of distinct arguments \( \lambda_1 > \ldots > \lambda_N \), no differentiation is required and the divided difference is a linear combination of the function values \( h(\lambda_1), \ldots, h(\lambda_N) \). Specifically, the recursion can be untwined to obtain [dB05, p. 60]

\[
\Delta(\lambda_1, \ldots, \lambda_N)h = \sum_{n=1}^{N} \frac{h(\lambda_n)}{\prod_{k \neq n}(\lambda_n - \lambda_k)}.
\]

(4.13)

In the more general case with the sequence \( \lambda_1 \geq \ldots \geq \lambda_N \) having \( M \leq N \) distinct values \( \mu_1 \geq \ldots \geq \mu_M \), with \( \mu_m \) occurring \( n_m \) times, the divided difference is instead a linear combination of the following function values and derivatives:

\[
\left\{ h^{(k_m)}(\mu_m) : m = 1, \ldots, M; \ k_m = 0, \ldots, n_m - 1 \right\}.
\]

The connection between divided differences and the generalized gamma distribution is suggested by comparing the two explicit formulas (4.11) and (4.13) for the case of distinct parameters \( \lambda_1 > \ldots > \lambda_N > 0 \). The following theorem details this connection in the general case allowing for repeated parameters.
Theorem 4.2. The PDF of the generalized gamma distribution with parameters \( \lambda_1 \geq \ldots \geq \lambda_N \) is given by the following divided difference:

\[
p_N^\Gamma(z; \lambda_1, \ldots, \lambda_N) = \Delta(\lambda_1, \ldots, \lambda_N)[\lambda^{N-2}e^{-z/\lambda}], \quad z \geq 0.
\]  

(4.14)

Proof. A formal proof is provided in Appendix 4.A.2, utilizing the contour-integral representation of the divided difference. The proof is inspired by that of [Kun65], which investigates the connection between divided differences and certain Laplace transforms.

We illustrate the usefulness of the divided-difference formulation to derive the following basic result on the PDF \( p_N^\Gamma(z; \lambda_1, \ldots, \lambda_N) \) around \( z = 0 \), which will be used in Section 4.3.

Lemma 4.1. The first non-zero derivative of \( p_N^\Gamma(z; \lambda_1, \ldots, \lambda_N) \) at \( z = 0 \) is given by

\[
\frac{\partial^{N-1}}{\partial z^{N-1}} p_N^\Gamma(z; \lambda_1, \ldots, \lambda_N) \bigg|_{z=0} = \frac{1}{\lambda_1 \ldots \lambda_N}.
\]

(4.15)

Proof. Due to the smoothness of \( \lambda^{N-2}e^{-z/\lambda} \) in (4.14), we can exchange the order of the differentiation and the divided difference operators to obtain

\[
\frac{\partial^k}{\partial z^k} p_N^\Gamma(z; \lambda_1, \ldots, \lambda_N) \bigg|_{z=0} = \Delta(\lambda_1, \ldots, \lambda_N)((-1)^k\lambda^{N-2-k})
\]

(4.16)

for \( k = 0, \ldots, N - 1 \). The proof is concluded by the following well-known result on divided differences of power functions [dB05]:

\[
\Delta(\lambda_1, \ldots, \lambda_N)(\lambda^i) = \begin{cases} 
0, & \text{if } i = 0, \ldots, N - 2, \\
\frac{(-1)^{N-1}}{\lambda_1 \ldots \lambda_N}, & \text{if } i = -1.
\end{cases}
\]

4Note that Lemma 4.1 is not novel and can be deduced from, e.g., [WG03].

4.2.2 Expressions for Average Performance

We next consider average performance on the form

\[
g_{\Gamma}(x) = \mathbb{E}[f(\tilde{\gamma}_\Gamma x)]
\]

with respect to the generalized gamma distribution with \( \tilde{\gamma}_\Gamma \sim \Gamma(\lambda_1, \ldots, \lambda_N) \). As mentioned earlier, this formulation is directly applicable for determining average performance for the last subchannel to be decoded under correlated Rayleigh fading in (4.6), with \( g_{\text{Ray}}^{\text{Ray}}(x) = \mathbb{E}[f(\tilde{\gamma}_{\text{Ray}} x)] \), since \( \tilde{\gamma}_{\text{Ray}} \sim \Gamma(\lambda_{R,1}, \ldots, \lambda_{R,n_R}) \). The following theorem provides a general method to compute average performance measures of this kind.
4.2. PERFORMANCE IN GENERALIZED GAMMA FADING

Theorem 4.3. For a given instantaneous performance measure \( f : \mathbb{R}_+ \to \mathbb{R} \), consider the average performances \( g_\Gamma(x) = \mathbb{E}[f(\tilde{\gamma}_\Gamma x)] \) and \( g_{\text{exp}}(x) = \mathbb{E}[f(\tilde{\gamma}_{\text{exp}} x)] \) with respect to normalized SNRs \( \tilde{\gamma}_\Gamma \sim \Gamma(\lambda_1, \ldots, \lambda_N) \) and \( \tilde{\gamma}_{\text{exp}} \sim \Gamma(1) \), where the latter follows a standard exponential distribution. \( g_\Gamma(x) \) and \( g_{\text{exp}}(x) \) are related via the divided difference

\[
g_\Gamma(x) = \Delta(\lambda_1, \ldots, \lambda_N) \left[ \lambda^{N-1} g_{\text{exp}}(\lambda x) \right]
\]

(4.17)

provided that \( g_\Gamma(x) \) is continuous and \( g_{\text{exp}}(x) \) is \((N-1)\) times continuously differentiable.

Proof. Let \( p_{\text{exp}}(z; \lambda) = e^{-z/\lambda}/\lambda \) denote the PDF of an exponential random variable with scale parameter \( \lambda \), and note that

\[
p^N_\Gamma(z; \lambda_1, \ldots, \lambda_N) = \Delta(\lambda_1, \ldots, \lambda_N)[\lambda^{N-1} p_{\text{exp}}(z; \lambda)].
\]

(4.18)

It follows for \( \lambda_1 > \ldots > \lambda_N \) that

\[
g_\Gamma(x) = \int_0^\infty f(zx) \Delta(\lambda_1, \ldots, \lambda_N)[\lambda^{N-1} p_{\text{exp}}(z; \lambda)]
\]

(4.19)

\[= \Delta(\lambda_1, \ldots, \lambda_N) \int_0^\infty f(zx) \lambda^{N-1} p_{\text{exp}}(z; \lambda)\]

\[= \Delta(\lambda_1, \ldots, \lambda_N) \left[ \lambda^{N-1} g_{\text{exp}}(\lambda x) \right].\]

The extension to the general case \( \lambda_1 \geq \ldots \geq \lambda_N \) follows by the continuity of the divided difference [HJ91, Thm. 6.1.26].

The theorem states that average performance under the generalized gamma distribution is completely characterized by the average performance under the exponential distribution. With distinct parameters, the average performance is simply given by the explicit formula (4.13) as

\[
g_\Gamma(x) = \sum_{n=1}^N \frac{\lambda_n^{N-1} g_{\text{exp}}(\lambda_n x)}{\prod_{k \neq n} (\lambda_n - \lambda_k)}.
\]

(4.20)

In the more general case with \( \lambda_1 \geq \ldots \geq \lambda_N \), the function \( g_\Gamma(x) \) can be obtained using the recursive procedure in Definition 4.1, which involves derivatives of \( \lambda^{N-1} g_{\text{exp}}(\lambda x) \) in \( \lambda \).

Average performance measures \( g_{\text{exp}}(x) = \mathbb{E}[f(\tilde{\gamma}_{\text{exp}} x)] \) under the exponential distribution is a well-studied matter in the literature.\(^5\) We have gathered explicit expressions for a sample of common performance measures in Table 4.1 in terms of elementary functions and the exponential

\(^5\)An exponentially distributed SNR is commonly referred to as “Rayleigh-fading”, but this terminology is not used here to avoid confusion with the MIMO Rayleigh-fading model.
Table 4.1: Performance measures with SNR $\gamma = \bar{\gamma}_{\text{exp}}x$ following an exponential distribution with scale parameter $x$.

<table>
<thead>
<tr>
<th>Perf. Measure</th>
<th>Notation</th>
<th>Definition</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average SNR</td>
<td>$g^{\text{SNR}}_{\text{exp}}(x)$</td>
<td>$\mathbb{E}[\gamma]$</td>
<td>$x$</td>
</tr>
<tr>
<td>Average MSE</td>
<td>$g^{\text{MSE}}_{\text{exp}}(x)$</td>
<td>$\mathbb{E}[1/\gamma]$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>Achiev. Rate</td>
<td>$g^{\text{Rate}}_{\text{exp}}(x)$</td>
<td>$\mathbb{E}[(\log(1 + \gamma)]$</td>
<td>$e^{1/x}E_1(1/x)$</td>
</tr>
<tr>
<td>Outage Prob.</td>
<td>$g^{\text{Outage}}<em>{\text{exp}}(x; \gamma</em>{\text{th}})$</td>
<td>$\Pr(\gamma \leq \gamma_{\text{th}})$</td>
<td>$1 - e^{-\gamma_{\text{th}}/x}$</td>
</tr>
<tr>
<td>Av. erfc$(\sqrt{\gamma})$</td>
<td>$g^{\text{erfc}}_{\text{exp}}(x)$</td>
<td>$\mathbb{E}[	ext{erfc}(\sqrt{\gamma})]$</td>
<td>$1 - \sqrt{\frac{1}{1+\frac{x}{\pi}}} \arctan \sqrt{\frac{1+x}{x}}$</td>
</tr>
<tr>
<td>Av. erfc$^2$(\sqrt{\gamma}))</td>
<td>$g^{\text{erfc}}_{\text{exp}}^2(x)$</td>
<td>$\mathbb{E}[	ext{erfc}^2(\sqrt{\gamma})]$</td>
<td>$1 - \sqrt{\frac{1}{1+\frac{x}{\pi}}}$</td>
</tr>
</tbody>
</table>

For composing average performance measures under the generalized gamma distribution, the case of average SNR serves as a trivial example. In this case we have $g^{\text{SNR}}_{\text{exp}}(x) = x$ and the divided difference formula in (4.17) reduces to

$$
g^{\text{SNR}}_{\Gamma}(x) = \Delta(\lambda_1, \ldots, \lambda_{n_R}) \left[ \lambda N \right] x,$$

where $g^{\text{erfc}}_{\Gamma}(x)$ and $g^{\text{erfc}}_{\Gamma}^2(x)$ are obtained using Theorem 4.3, and the constants $\theta$, $\beta$, and $\delta_n$ were defined in Section 3.2.2.

Interestingly, performance in terms of the MSE requires special attention. The reason is that Theorem 4.3 does not apply, since the average MSE with an exponentially distributed SNR is infinite. However, the average

$$\int_1^\infty e^{-y/t}t \, dt.$$

These are obtained using the list of performance measures previously considered in Section 3.2.2.

This is closely connected with the fact that the average MSE is infinite for the subchannel decoded first in a fully multiplexed systems with $n_T \geq n_R = n_L$ using a ZF-DF receiver and the statistical channel models under consideration. This does, however, not mean that the subchannel is useless: Average bit and symbol error probabilities still tend to zero with increasing average SNR.
MSE is finite for a generalized-gamma distributed normalized SNR of order $N \geq 2$. Therefore, it is necessary to develop another rule for determining the average MSE. As we show next, the finite average MSE can be stated in terms of another divided difference.

**Theorem 4.4.** The average MSE under the generalized gamma distribution is given by

$$g_{\text{MSE}}^\Gamma(x) = \frac{\Delta(\lambda_1, \ldots, \lambda_N) \left[ \lambda^{N-2} \log(\lambda) \right]}{x}, \quad N \geq 2.$$  

(4.22)

**Proof.** The proof is located in Appendix 4.A.3. \qed

Note that $g_{\text{MSE}}^\Gamma(x)$ can also be expressed on the form in (4.17). By defining $\tilde{g}_{\text{MSE}}^\exp(x) = \log(x)/x$, (4.22) is easily manipulated to obtain

$$g_{\text{MSE}}^\Gamma(x) = \Delta(\lambda_1, \ldots, \lambda_N) \left[ \lambda^{N-1} \tilde{g}_{\text{MSE}}^\exp(\lambda x) \right], \quad N \geq 2.$$  

(4.23)

In a completely analogous manner, we can obtain expressions for average performance under a generalized gamma product distribution on the form

$$\tilde{\gamma}_{2 \times \Gamma} \sim \tilde{\gamma}_{\Gamma, 1} \cdot \tilde{\gamma}_{\Gamma, 2},$$  

(4.24)

where $\tilde{\gamma}_{\Gamma, 1} \sim \Gamma(\lambda_1, 1, \ldots, \lambda_{1,N_1})$ and $\tilde{\gamma}_{\Gamma, 2} \sim \Gamma(\lambda_2, 1, \ldots, \lambda_{2,N_2})$ are statistically independent. This case applies to the last subchannel under the double-scattering model, for which $\tilde{\gamma}_{\text{mL}}^\text{DS} \sim z_R^H R_R z_R \cdot z_S^H R_S z_S$. It is easily verified that

$$g_{\text{SNR}}^{2 \times \Gamma}(x) = (\lambda_1 + \ldots + \lambda_{1,N_1})(\lambda_2 + \ldots + \lambda_{2,N_2}) \cdot x,$$  

(4.25)

$$g_{\text{MSE}}^{2 \times \Gamma}(x) = \Delta(\lambda_1, \ldots, \lambda_{1,N_1}) \left[ \lambda_1^{N_1-2} \log(\lambda_1) \right],$$

$$\Delta(\lambda_2, \ldots, \lambda_{2,N_2}) \frac{\lambda_2^{N_2-2} \log(\lambda_2)}{x}.$$  

Expressions for other performance measures can be obtained using the following theorem.

**Theorem 4.5.** For a given instantaneous performance measure $f : \mathbb{R}_+ \rightarrow \mathbb{R}$, consider the average performances $g_{2 \times \Gamma}(x) = \mathbb{E}[f(\tilde{\gamma}_{2 \times \Gamma} x)]$ and $g_{2 \times \exp}(x) = \mathbb{E}[f(\tilde{\gamma}_{2 \times \exp} x)]$ with respect to normalized SNRs $\tilde{\gamma}_{2 \times \Gamma}$ in (4.24) and $\tilde{\gamma}_{2 \times \exp} \sim \tilde{\gamma}_{\exp, 1} \cdot \tilde{\gamma}_{\exp, 2}$, where the latter is a product of two independent standard exponential random variables. $g_{2 \times \Gamma}(x)$ and $g_{2 \times \exp}(x)$ are related via the following divided difference in two variables

$$g_{2 \times \Gamma}(x) = \Delta(\lambda_1, \ldots, \lambda_{1,N_1}) \left[ \Delta(\lambda_2, \ldots, \lambda_{2,N_2}) \left[ g(\lambda_1, \lambda_2; x) \right] \right],$$
74 CHAPTER 4. PERFORMANCE IN CORRELATED CHANNELS

Table 4.2: Performance measures with SNR $\gamma = \bar{\gamma}_{\text{exp},1} \bar{\gamma}_{\text{exp},2} x$ following a product-exponential distribution with scale parameter $x$.

<table>
<thead>
<tr>
<th>Perf. Measure</th>
<th>Notation</th>
<th>Definition</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average SNR</td>
<td>$g_{\text{SNR}}^{2\times\exp}(x)$</td>
<td>$\mathbb{E}[\gamma]$</td>
<td>$x$</td>
</tr>
<tr>
<td>Average MSE</td>
<td>$g_{\text{MSE}}^{2\times\exp}(x)$</td>
<td>$\mathbb{E}[1/\gamma]$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>Achiev. Rate</td>
<td>$g_{\text{Rate}}^{2\times\exp}(x)$</td>
<td>$\mathbb{E}[\log(1 + \gamma)]$</td>
<td>$G_{2,4}^{4,1}\left(\frac{1}{x}, 0, 0, 0, 0\right)$</td>
</tr>
<tr>
<td>Outage Prob.</td>
<td>$g_{\text{Outage}}^{2\times\exp}(x; \gamma^{th})$</td>
<td>$\Pr(\gamma \leq \gamma^{th})$</td>
<td>$1 - \sqrt{\frac{\gamma^{th}}{x}} K_0\left(\sqrt{\frac{\gamma^{th}}{x}}\right)$</td>
</tr>
<tr>
<td>Av. erfc($\sqrt{\gamma}$)</td>
<td>$g_{\text{erfc}}^{2\times\exp}(x)$</td>
<td>$\mathbb{E}[\text{erfc}(\sqrt{\gamma})]$</td>
<td>$\frac{1}{\sqrt{\pi}} G_{3,2}^{2,2}\left(\frac{1}{x}, \frac{1}{\sqrt{2}}\right)$</td>
</tr>
</tbody>
</table>

where

$$g(\lambda_1, \lambda_2; x) = \lambda_1^{N_1-1} \lambda_2^{N_2-1} g_{2\times\exp}(\lambda_1 \lambda_2 x),$$

provided that $g_{2\times\exp}(x)$ is continuous and $g_{2\times\exp}(x)$ is $(N_1 + N_2 - 2)$ times continuously differentiable.

Proof. The proof is very similar to that of Theorem 4.3 and is therefore omitted.

For convenience, we have gathered a number of relevant performance measures in Table 4.2 for a normalized SNR being the product of two independent standard exponential random variables. These are expressed in terms of the modified Bessel function of the second kind, $K_0(x)$, as well as the Meijer G function. These expressions can be found in [SL03a, SL03b, MS06]. The contribution in Theorem 4.5 is an extension to the more general case where the normalized SNR is instead a product of two independent generalized-gamma random variables.

4.3 SNR Approximations for ZF Receivers

In this section, we derive approximate distributions for the marginal normalized SNRs $\bar{\gamma}_1^{\text{Ray}}, \ldots, \bar{\gamma}_{n_l-1}^{\text{Ray}}$ and $\bar{\gamma}_1^{\text{DS}}, \ldots, \bar{\gamma}_{n_l-1}^{\text{DS}}$. The approximations are based on the generalized gamma and gamma-product distributions, respectively. In light of the preceding section, this is a convenient choice since the techniques derived therein then apply to form expressions for performance measures on all subchannels. This choice is also natural, since each $\bar{\gamma}_n^{\text{Ray}}$ and $\bar{\gamma}_n^{\text{DS}}$ can be represented as a statistical mixture of generalized gamma and gamma-product distributions, respectively (see (4.43) in Appendix 4.A.4).
4.3. SNR APPROXIMATIONS FOR ZF RECEIVERS

Asserting that $\bar{\gamma}_{\text{Ray}}$ or $\bar{\gamma}_{\text{DS}}$ is to be approximated by a member in a parameterized family of distributions leads naturally into classical estimation. By drawing independent samples from the distribution it is possible to, for example, match moments or use a maximum-likelihood estimate [Kay93]. Here we instead choose an analytical approach that does not rely on estimation. The normalized SNR approximations presented in this section have the following properties by design: First, the approximate PDF $\hat{p}_n(\bar{\gamma})$ is fitted to the true PDF $p_n(\bar{\gamma})$ as $\bar{\gamma} \to 0$. By a result in [WG03], the first non-zero derivative of $p_n(\bar{\gamma})$ at $\bar{\gamma} = 0$ dictates the behavior of average error probabilities as well as the outage probability in the high “average SNR” regime (as $x \to \infty$). Second, by selecting the approximation from a suitable family of distributions, sharing main characteristics with the true PDF, we may expect that the approximations perform well for medium to low average SNRs as well.

In order to present the approximations, we shall derive a few results that require the following auxiliary distribution.

**Definition 4.3 (Correlated Bartlett Distribution).** Given $\lambda_1 \geq \ldots \geq \lambda_N > 0$ and $\mathbf{Z} \in \mathbb{C}^{N \times N}$ with i.i.d. $\mathcal{CN}(0, 1)$ entries, let

$$\mathbf{Q} \mathbf{L} = \text{diag} \left( \sqrt{\lambda_1}, \ldots, \sqrt{\lambda_N} \right) \mathbf{Z}$$

denote a QL-decomposition with a unitary $\mathbf{Q} \in \mathbb{C}^{N \times N}$ and a lower triangular $\mathbf{L} \in \mathbb{C}^{N \times N}$ having non-negative (real) diagonal entries. Then $\mathbf{L}$ is said to follow a correlated Bartlett distribution of order $N$ with parameters $\lambda_1, \ldots, \lambda_N$, denoted as $\mathbf{L} \sim \mathcal{B}(\lambda_1, \ldots, \lambda_N)$.

We shall state a lemma that characterizes the squared diagonal elements of $\mathbf{L} \sim \mathcal{B}(\lambda_1, \ldots, \lambda_N)$. Note that an equivalent result has previously been presented in [KS04a]. For completeness, however, we include a novel proof that makes use of the generalized gamma distribution and divided differences. In the following, $e_m(\lambda_1, \ldots, \lambda_N)$ denotes the $m^{\text{th}}$ elementary symmetric polynomial in $N$ variables:

$$e_m(\lambda_1, \ldots, \lambda_N) = \sum_{1 \leq j_1 < \ldots < j_m \leq N} \lambda_{j_1} \ldots \lambda_{j_m}.$$

Two important special cases of elementary symmetric polynomials are

$$e_m(\lambda_1, \ldots, \lambda_N) = \begin{cases} \sum_{n=1}^{N} \lambda_n, & \text{if } m = 1, \\ \prod_{n=1}^{N} \lambda_n, & \text{if } m = N. \end{cases}$$

---

8The name used for this distribution here is inspired by the Bartlett decomposition in statistics [Mui82], which is the triangular Cholesky decomposition of a Wishart matrix.
Lemma 4.2. Let $L \sim \mathcal{B}(\lambda_1, \ldots, \lambda_N)$, and let $p_n(z)$ denote the PDF of $l^2_{nn}$, and let $p^{(k)}_n(z)$ denote the $k$th derivative of $p_n(z)$. The first non-zero derivative of $p_n(z)$ at $z=0$ is

$$p^{(n-1)}_n(0) = \binom{N}{n}^{-1} \frac{e_{N-n}(\lambda_1, \ldots, \lambda_N)}{e_N(\lambda_1, \ldots, \lambda_N)}.$$ 

Proof. The proof is located in Appendix 4.A.4. \[ \Box \]

We are now ready to state the main result of this section. Given $L \sim \mathcal{B}(\lambda_1, \ldots, \lambda_N)$, it shows how the leading $n \times n$ submatrix of $L$ can be approximated by a correlated Bartlett matrix $\hat{L}^{(n)} \sim \mathcal{B}(\lambda^{(n)}_1, \ldots, \lambda^{(n)}_n)$ of order $n$.

Theorem 4.6. Consider $L \sim \mathcal{B}(\lambda_1, \ldots, \lambda_N)$, with squared diagonal elements $l^2_{11}, \ldots, l^2_{NN}$ and associated PDF:s $p_1(x), \ldots, p_N(x)$. Similarly, given $n$ in $1 \leq n \leq N - 1$, consider $\hat{L}^{(n)} \sim \mathcal{B}(\lambda^{(n)}_1, \ldots, \lambda^{(n)}_n)$, with squared diagonal elements $\hat{l}^2_{11}, \ldots, \hat{l}^2_{nn}$ and PDF:s $\hat{p}_1(x), \ldots, \hat{p}_n(x)$. Given $\lambda_1 \geq \ldots \geq \lambda_N > 0$, there exists a unique sequence $\lambda^{(n)}_1 \geq \ldots \geq \lambda^{(n)}_n > 0$ such that

$$p^{(k)}_m(x) = \hat{p}^{(k)}_m(x), \quad \begin{cases} m = 1, \ldots, n, \\ k = 0, \ldots, m - 1. \end{cases}$$

where $p^{(k)}_m(x)$ and $\hat{p}^{(k)}_m(x)$ denote the $k$th derivative of $p_m(x)$ and $\hat{p}_m(x)$, respectively. The sequence $\lambda^{(n)}_1 \geq \ldots \geq \lambda^{(n)}_n > 0$ is identified as the $n$ real roots of the following polynomial equation in $t$:

$$\sum_{m=0}^{n} (-1)^{n-m} \binom{N-m}{N-n} \frac{e_{N-m}(\lambda_1, \ldots, \lambda_N)}{e_{N-n}(\lambda_1, \ldots, \lambda_N)} t^m = 0. \quad (4.27)$$

Proof. The proof is located in Appendix 4.A.5 \[ \Box \]

Part of the theorem states that the polynomial equation in (4.27) is guaranteed to have $n$ real, positive roots (counting multiplicities), which is necessary for the application of the theorem. The roots can be found efficiently by numerical means. For a given $L \sim \mathcal{B}(\lambda_1, \ldots, \lambda_N)$, each $\hat{L}^{(n)} \sim \mathcal{B}(\lambda^{(n)}_1, \ldots, \lambda^{(n)}_n)$ obtained in Theorem 4.6 provides a generalized gamma approximation $\hat{l}^2_{nn} \sim \Gamma(\lambda^{(n)}_1, \ldots, \lambda^{(n)}_n)$ of $l^2_{nn}$. The random variable $\hat{l}^2_{nn}$ is identified as the $n$th squared diagonal element of $\hat{L}^{(n)}$. This approach directly applies to finding generalized gamma approximations for normalized SNRs $\hat{\gamma}^{\text{Ray}}_1, \ldots, \hat{\gamma}^{\text{Ray}}_{n-1}$ under the correlated Rayleigh channel model. In terms of a correlated Bartlett matrix $L_R \sim \mathcal{B}(\lambda_{R,1}, \ldots, \lambda_{R,n_R})$ in the eigenvalues of $R_R$, Theorem 4.1 states that the normalized SNRs
are distributed as $\tilde{\gamma}_n^{Ray} \sim l_{R,kk}^2$, where $k = n_R - n_L + n$. For the double-scattering model, we additionally make use of a correlated Bartlett matrix $L_S \sim B(\lambda_{S,1}, \ldots, \lambda_{S,n_S})$ in the eigenvalues of $R_S$. The normalized SNRs are given by $\hat{\gamma}_n^{Ray} \sim l_{R,k_1}^2 l_{S,k_2}^2$, where $k_1 = n_R - n_L + n$ and $k_2 = n_S - n_L + n$, and Theorem 4.6 can be invoked separately to obtain generalized gamma approximations for the two factors.

4.4 Numerical Results

We shall provide a few examples that illustrate the accuracy of the proposed approximations. We consider a $4 \times 4$ MIMO system with a ZF-DF receiver under the assumption of no error propagation. Full spatial multiplexing with 4 subchannels is assumed, and the proposed expressions for average performance in this chapter are compared with their Monte-Carlo estimated true counterparts. Note that the proposed expressions are asserted to be exact for the last subchannel to be decoded. The normalized SNR vectors $\tilde{\gamma}^{Ray}$ and $\tilde{\gamma}^{DS}$ are characterized by Theorem 4.1 as

$$\tilde{\gamma}^{Ray} \sim l^2(\bar{Z}^H R_R \bar{Z}),$$
$$\tilde{\gamma}^{DS} \sim l^2(\bar{Z}_R^H R_R \bar{Z}_R) \odot l^2(\bar{Z}_S^H R_S \bar{Z}_S),$$

where $\bar{Z}$, $\bar{Z}_R$, and $\bar{Z}_S$ have i.i.d. $CN(0,1)$ entries. The receive and scattering correlation matrices $R_R \in \mathbb{C}^{4 \times 4}$ and $R_S \in \mathbb{C}^{4 \times 4}$ are both generated using the exponential correlation model described in Section 3.4: the real correlation parameters $\rho_R$ and $\rho_S$ control the “amount” of receive and scattering correlation present, respectively, in the range $\rho_R, \rho_S \in [0,1]$.

Figures 4.1 and 4.2 display the proposed expressions for the average BEP using square QAM constellations under the two channel models. The true values are estimated using $10^7$ channel realizations, and instantaneous performance is computed using the exact expression in (3.17). It is observed that the approximate expressions for subchannels 1–3 accurately reflect the true average BEP. While this is guaranteed for high “average SNRs” by design of the approximations, it is verified that these expressions are in fact accurate over the entire SNR range.

In Figure 4.3 we instead examine the average MSEs of the subchannels in the case of a correlated Rayleigh model. Due to the reciprocal relation between the MSE and the SNR for ZF receivers in (3.10), the true average MSEs are on the form

$$E \left[ (\hat{\gamma}_n^{Ray})^{-1} \right] = E \left[ (\tilde{\gamma}_n^{Ray})^{-1} \right] x^{-1}, \quad (4.28)$$

for subchannels 2–4, while the average MSE is infinite for subchannel 1. The proposed expressions for the average MSE in Theorem 4.4 are on the
Figure 4.1: Proposed expressions (solids) and estimated true values (markers) of average BEP using 16-QAM in correlated Rayleigh fading with $\rho_R = 0.7$.

Figure 4.2: Proposed expressions (solids) and estimated true values (markers) of average BEP using 4-QAM in double-scattering fading with $\rho_R = \rho_S = 0.7$. 
4.5. SUMMARY

In this chapter we have presented analytic expressions for a range of average subchannel performance measures for normalized SNRs following the generalized gamma and gamma-product distributions. These directly apply to evaluate performance for the last subchannel in line to be decoded under the two statistical channel models under consideration: the separately-correlated Rayleigh and double-scattering models. For the other subchannels we have presented means to approximate the true normalized SNRs using generalized gamma distributions leading to approximate expressions

$$g_n^{\text{MSE}}(x) = g_n^{\text{MSE}} x^{-1}, \quad n \geq 2,$$

for a constant $g_n^{\text{MSE}}$. Since there will be a constant ratio between the expressions in (4.28) and (4.29) for different values of the scale parameter $x$, we have instead compared the true and approximate constants, or normalized MSEs, over the range of the receive correlation parameter. It is verified using $10^6$ channel realizations that the approximate expressions for subchannels 2 and 3 are quite accurate, especially for low values of the correlation parameter. This is expected, since the approximate distributions coincide with the true distributions for uncorrelated channels ($\rho_R = 0$). The negative effect of receive correlation on performance in terms of the average MSE is also evident in this example.
for average performance. We also concluded using Theorem 4.1 that such performance measures are on the very same form as those considered in the previous chapter. This insight extends the optimization framework for MIMO transceiver design to include the correlated MIMO channel models considered in this chapter.
4.A  A Collection of Proofs

4.A.1 Proof of Theorem 4.1

The proof relies on Lemma 3.4, which is reiterated here without proof for convenience.

Lemma 4.3. For a positive semidefinite matrix $T \in \mathbb{C}^{M \times M}$ and a tall matrix $S \in \mathbb{C}^{M \times N}$ having a QL decomposition $QL = S$,

$$l^2(S^H TS) = l^2(Q^H TQ) \odot l^2(S^H S).$$

Consider first the correlated Rayleigh-fading model. With $\gamma = l^2(P^H H^H H^T P)$ in (3.7) it follows from Lemma 4.3 that

$$\gamma = l^2(P^H R^1/2 Z^H R Z R^1/2 P) \odot l^2(P^H R^T P),$$

where the semi-unitary matrix $Q \in \mathbb{C}^{n_T \times n_L}$ stems from the QL decomposition $QL = R^1/2 S$. It follows from the unitary invariance of $Z$ that $ZQ \sim \bar{Z}$ irrespective of $Q$, and hence that $l^2(Q^H Z^H R Z Q) \sim l^2(\bar{Z}^H R \bar{Z})$, which implies the desired result in (4.2).

The double-scattering model can be similarly treated. Invoking Lemma 4.3 again, we get

$$\gamma \sim l^2(Q^H Z^H R Z R Q) \odot l^2(P^H R^1/2 Z^H S R Z S R^1/2 P),$$

where $Q \in \mathbb{C}^{n_T \times n_L}$ is now a random semi-unitary matrix from the QL decomposition $QL = \bar{R}^{1/2} S$. Nevertheless, the two Cholesky vectors in (4.31) are independent since $Z_R$ and $Z_S$ are independent. Indeed, conditioned on $Z_S$, $Q$ is fixed and $Z_R Q \sim Z_R$. The conditional distribution does not depend on $Z_S$, and it follows that $Z_S$ and $Z_R Q$ are independent. Proceeding with the second factor in (4.31) as in (4.30) leads to the desired result.

4.A.2 Proof of Theorem 4.2

Let $\tilde{\gamma} \sim \lambda_1 X_1 + \ldots + \lambda_N X_N \sim \Gamma(\lambda_1, \ldots, \lambda_N)$ according to Definition 4.1. Allowing for repeated values among $\lambda_1, \ldots, \lambda_N > 0$, let there be $M$ distinct values $\mu_1, \ldots, \mu_M$, with $\mu_m$ occurring $n_m$ times ($m = 1, \ldots, M$). With the aid of the characteristic function,

$$\int_{-\infty}^{\infty} p_N^\Gamma(z; \lambda_1, \ldots, \lambda_N)e^{iz}dz = \mathbb{E}[e^{i\tilde{\gamma}}] = \mathbb{E}[e^{i\lambda_1 X_1}] \ldots \mathbb{E}[e^{i\lambda_N X_N}] = (1 - it\mu_1)^{-n_1} \ldots (1 - it\mu_M)^{-n_M}.$$
The result follows if
\[
I(t) = \int_0^\infty \Delta(\lambda_1, \ldots, \lambda_N)[\lambda^{N-2}e^{-z/\lambda}]e^{itz}dz
\] (4.33)

coincides with (4.32). For a closed contour \(C\) enclosing \(\mu_1, \ldots, \mu_M\) on \(\text{Re}(\xi) > 0\), we have [Kum65, Mil00]
\[
\Delta(\lambda_1, \ldots, \lambda_N)[\lambda^{N-2}e^{-z/\lambda}] = \frac{1}{2\pi i} \oint_C \frac{\xi^{N-2}e^{-z/\xi}}{\xi^{n_1}\cdots(\xi - \mu_M)^{n_M}}d\xi,
\] (4.34)

since \(\xi^{N-2}e^{-z/\xi}\) is holomorphic on \(\text{Re}(\xi) > 0\). By exchanging the order of integration,
\[
I(t) = \frac{1}{2\pi i} \oint_C \frac{\xi^{N-2}}{\xi^{n_1}\cdots(\xi - \mu_M)^{n_M}} \int_0^\infty e^{-z/\xi}e^{itz}dzd\xi
\]
\[
= -\frac{1}{it} \frac{1}{2\pi i} \oint_C \frac{\xi^{N-1}}{\xi^{n_1}\cdots(\xi - \mu_M)^{n_M}(\xi + it^{-1})}d\xi.
\]

Fix \(t\), and let \(C'\) enclose the interior of \(C\) as well as \(-it^{-1}\), and let \(C''\) enclose only \(-it^{-1}\). Using \(\oint_C = \oint_{C'} - \oint_{C''}\), and the contour representation of the divided difference\(^9\), we obtain
\[
I(t) = -\frac{1}{it} \Delta(-it^{-1}, \lambda_1, \ldots, \lambda_N)[\lambda^{N-1}]
\] (4.35)
\[
+ \frac{1}{it} \frac{1}{2\pi i} \oint_{C''} \frac{\xi^{N-1}}{\xi^{n_1}\cdots(\xi - \mu_M)^{n_M}(\xi + it^{-1})}d\xi.
\]

The first term vanishes, since an order-\(N\) divided difference of a polynomial with degree less than \(N\) is zero [dB05]. The second term is given by the residue as
\[
I(t) = \frac{1}{it} \frac{(it)^{-(N-1)}}{(-it^{-1} - \mu_1)^{n_1}\cdots(-it^{-1} - \mu_M)^{n_M}},
\] (4.36)
which is easily seen to coincide with (4.32).

4.A.3 Proof of Theorem 4.4

Note first that
\[
\frac{d^{N-1}}{d\lambda^{N-1}}[\lambda^{N-2}\log(\lambda)] = \frac{(N - 2)!}{\lambda},
\] (4.37)
\(^9\)This definition holds for complex arguments as well.
Proof. From the QL-decomposition $\bar{Q}L = D_{\sqrt{\lambda}}$ it follows that $\text{span}(D_{\sqrt{\lambda}}Q_2) = \text{span}(\bar{Q}_2)$. Since $Q = \bar{Q}$, we observe that $Q_2^H D_{\sqrt{\lambda}}Q_1 = 0$, and furthermore that

$$Q_1^H D_{\sqrt{\lambda}}Q_1 = Q_1^H D_{\sqrt{\lambda}} (I - \bar{Q}_2 \bar{Q}_2^H) D_{\sqrt{\lambda}} Q_1,$$

$$Q_1^H D_{\sqrt{\lambda}} Q_1 \bar{Q}_1^H D_{\sqrt{\lambda}} Q_1.$$
Taking determinants and using \( \det(AB) = \det(BA) \) for square \( A \) and \( B \) results in
\[
\det \left( Q_1^H D_{\lambda} Q_1 \right) = \det \left( \bar{Q}_1^H D_{\sqrt{\lambda}} Q_1 Q_1^H D_{\sqrt{\lambda}} \bar{Q}_1 \right) = \det \left( \bar{L}_1^H \bar{L}_1 \right),
\]
where \( \bar{L}_1 = \bar{Q}_1^H D_{\sqrt{\lambda}} Q_1 \) and \( Q_1 = \bar{Q}_1 \), and the result follows. \( \square \)

We now prove the desired result in Lemma 4.2. \( l_{nn}^2 \) can be expressed as
\[
l_{nn}^2 = \bar{z}_n^H D_{\sqrt{\lambda}} (I - Q_2 Q_2^H) D_{\sqrt{\lambda}} \bar{z}_n,
\]
where \( \bar{z}_n \) denotes the \( n \)th column of \( \bar{Z} \). Conditioned on \( Q_2 \), \( l_{nn}^2 \) follows a generalized gamma distribution with parameters \( \lambda_1^{[n]} \geq \ldots \geq \lambda_n^{[n]} \) being the \( n \) largest eigenvalues of \( D_{\sqrt{\lambda}} (I - Q_2 Q_2^H) D_{\sqrt{\lambda}} \). These are precisely the eigenvalues of \( Q_1^H D_{\lambda} Q_1 \), and it follows by Lemma 4.1 that the first non-zero derivative of \( p_n(x) \) at \( x = 0 \) is
\[
p_n^{(n-1)}(0) = \mathbb{E} \left[ \frac{1}{\det(\bar{Q}_1^H D_{\lambda} Q_1)} \right].
\]
This expectation can be reformulated as follows. Taking determinants of \( \bar{L}^H \bar{L} = Q^H D_{\lambda} Q \) we get
\[
\det(\bar{L}_1^H \bar{L}_1) \det(\bar{L}_2^H \bar{L}_2) = \prod_{k=1}^{N} \lambda_k.
\]
Since \( \bar{L}_2^H \bar{L}_2 = \det(\bar{Q}_2^H D_{\lambda} \bar{Q}_2) \), we conclude using Lemma 4.4 that
\[
p_n^{(n-1)}(0) = \mathbb{E} \left[ \frac{\det(\bar{Q}_2^H D_{\lambda} \bar{Q}_2)}{\prod_{k=1}^{N} \lambda_k} \right].
\]
Observing that \( \bar{Q}_2 \in \mathbb{C}^{N \times (N-n)} \) is uniformly distributed on the complex Stiefel manifold of semi-unitary matrices, the proof is concluded by the following lemma.

**Lemma 4.5.** Let \( Q \in \mathbb{C}^{N \times M} (M \leq N) \) be uniformly distributed on the Stiefel manifold of semi-unitary matrices. Then
\[
\mathbb{E} \left[ \det(\bar{Q}^H \text{diag}(\lambda_1, \ldots, \lambda_N) Q) \right] = \binom{N}{M}^{-1} e_M(\lambda_1, \ldots, \lambda_N).
\]
Proof. First we prove an auxiliary result. Let $S^M$ denote the family of $M$-subsets of $\{1, \ldots, N\}$, and given $S \in S^M$ let $Q_S$ denote the submatrix of $Q$ obtained by excluding any row with index $i \notin S$. By the Cauchy-Binet formula,

$$1 = \det (Q^H Q) = \sum_{S' \in S^M} \det (Q^H_{S'} Q_{S'}). \quad (4.47)$$

Since $Q$ is uniformly distributed, the matrices $\{Q_{S'}\}_{S' \in S^M}$ have the same marginal distribution. Taking expectations of (4.47) then gives, for any $S \in S^M$,

$$\mathbb{E} \left[ \det (Q^H_S Q_S) \right] = \binom{N}{M}^{-1}, \quad (4.48)$$

since $S^M$ has $\binom{N}{M}$ elements. Similarly, it follows by the Cauchy-Binet formula that

$$\det (Q^H \text{diag} (\lambda_1, \ldots, \lambda_N) Q) = \sum_{S \in S^M} \det (Q^H_S D_{SS} Q_S), \quad (4.49)$$

where $D_{SS}$ is obtained from $\text{diag} (\lambda_1, \ldots, \lambda_N)$ by excluding any row and column with index $i \notin S$. Taking expectations and using (4.48), we get

$$\mathbb{E} \left[ \det (Q^H \text{diag} (\lambda_1, \ldots, \lambda_N) Q) \right] = \sum_{S \in S^M} \mathbb{E} \left[ \det (Q^H_S Q_S) \right] \det (D_{SS})$$

$$= \binom{N}{M}^{-1} \sum_{1 \leq j_1 < \ldots < j_m \leq N} \lambda_{j_1} \cdots \lambda_{j_m},$$

which concludes the proof. \qed

4.A.5 Proof of Theorem 4.6

It follows by Lemma 4.2 that, irrespective of the choice of $\lambda_1^{(n)} \geq \ldots \geq \lambda_n^{(n)} > 0$,

$$p_m^{(k)}(0) = \hat{p}_m^{(k)}(0) = 0, \quad \begin{cases} m = 1, \ldots, n, \\ k = 0, \ldots, m - 2, \end{cases}$$

Ensuring that also

$$p_m^{(m-1)}(0) = \hat{p}_m^{(m-1)}(0), \quad m = 1, \ldots, n,$$
amounts to setting, for each \( m = 1, \ldots, n, \)
\[
\left( \begin{array}{c}
N \\
m
\end{array} \right)^{-1} \frac{e_{N-m}(\lambda_1, \ldots, \lambda_N)}{e_N(\lambda_1, \ldots, \lambda_N)} = \left( \begin{array}{c}
n \\
m
\end{array} \right)^{-1} \frac{e_{n-m}(\lambda_1^{(n)}, \ldots, \lambda_n^{(n)})}{e_n(\lambda_1^{(n)}, \ldots, \lambda_n^{(n)})}.
\]

Reformulating these equations slightly, we obtain
\[
e_{n-m}(\lambda_1^{(n)}, \ldots, \lambda_n^{(n)}) = \left( \begin{array}{c}
N - m \\
n - n
\end{array} \right) \frac{e_{N-m}(\lambda_1, \ldots, \lambda_N)}{e_{n-m}(\lambda_1, \ldots, \lambda_N)}, \quad m = 1, \ldots, n.
\]

(4.50)

The elementary symmetric polynomials are related to their arguments in the sense that \( \lambda_1^{(n)}, \ldots, \lambda_n^{(n)} \) are the roots of
\[
\sum_{m=0}^{n} (-1)^{n-m} e_{n-m}(\lambda_1^{(n)}, \ldots, \lambda_n^{(n)}) t^m = 0, \tag{4.51}
\]
which together with (4.50) amounts to the desired result in (4.27).

The proof is complete if we can show that the polynomial equation in (4.27) has \( n \) real roots counting multiplicities. Since \( \lambda_1, \ldots, \lambda_N \) are the roots of
\[
\sum_{m=0}^{N} (-1)^{N-m} e_{N-m}(\lambda_1, \ldots, \lambda_N) t^m = 0, \tag{4.52}
\]
the reciprocals \( \lambda_1^{-1}, \ldots, \lambda_N^{-1} \) are the roots of
\[
\sum_{m=0}^{N} (-1)^{N-m} e_{N-m}(\lambda_1, \ldots, \lambda_N) t^{N-m} = 0. \tag{4.53}
\]
The \((N - n)\)th derivative of the polynomial in (4.53) then has \( n \) zeroes \( \tilde{\lambda}_1, \ldots, \tilde{\lambda}_n \) (counting multiplicities) satisfying \( \tilde{\lambda}_N^{-1} \geq \tilde{\lambda}_1 \geq \ldots \geq \tilde{\lambda}_n \geq \lambda_1^{-1} \). Carrying out the differentiation results in \( \tilde{\lambda}_1, \ldots, \tilde{\lambda}_n \) being the roots of
\[
\sum_{m=0}^{n} (-1)^{n-m} \left( \begin{array}{c}
N - m \\
n - n
\end{array} \right) \frac{e_{N-m}(\lambda_1, \ldots, \lambda_N)}{e_{n-m}(\lambda_1, \ldots, \lambda_N)} t^{n-m} = 0. \tag{4.54}
\]
It follows that \( \tilde{\lambda}_1^{-1}, \ldots, \tilde{\lambda}_n^{-1} \) are the \( n \) real roots of (4.27), counting multiplicities.
Chapter 5

Optimization with Majorization Constraints

The contributions to MIMO transceiver design in Chapter 3 revolved around posing convex optimization problems that can be efficiently implemented using standard algorithms. This chapter takes a closer look at the underlying structure of certain optimization problems arising in MIMO transceiver design. Sections 5.1 and 5.2 investigate two different problem formulations related to MIMO transceiver design, and establish connections to convex-hull algorithms and optimization with respect to submodular constraints. The usefulness of these findings is exemplified in Section 5.3 by solving a precoder-optimization problem with strict per-antenna power constraints and heterogeneous constellations on the subchannels. It is also shown that—for a class of MSE-based cost functions—the statistical precoding problem in Chapter 3 can be solved in $O(n_L \log n_L)$ time using a simple convex-hull algorithm.

5.1 Optimization with Skewed Majorization Constraints

A central theme in this section is the interplay between the two closely related concepts: majorization of vectors and Schur-convexity of functions. These concepts were introduced in Chapter 3, but are reiterated here for clarity:

*Definition 5.1.* For $x, y \in \mathbb{R}^N$, $x$ is said to be majorized by $y$, or $x \preceq y$, if...
and only if
\[
\sum_{k=1}^{n} x_{\downarrow k} \leq \sum_{k=1}^{n} y_{\downarrow k}, \quad n = 1, \ldots, N - 1, \tag{5.1}
\]
\[
\sum_{k=1}^{N} x_k = \sum_{k=1}^{N} y_k, \tag{5.2}
\]
where \(x_{\downarrow k}\) and \(y_{\downarrow k}\) denote the \(k\)th largest components of \(x\) and \(y\), respectively. A real-valued function \(\phi\) defined on a set \(A \subset \mathbb{R}^N\) is said to be Schur-convex on \(A\), if \(x \preceq y\) on \(A\) implies \(\phi(x) \leq \phi(y)\).

Consider minimizing an objective function \(F : \mathbb{R}^N \to \mathbb{R}\) over a set defined by a majorization relation as
\[
\begin{align*}
\text{minimize} & \quad F(z) \\
\text{subject to} & \quad z + m \preceq y,
\end{align*}
\tag{5.3}
\]
where \(m\) and \(y\) are two given \(\mathbb{R}^N\)-vectors. Due to the asymmetry that is introduced by the vector \(m\), we refer to the constraint as a skewed majorization constraint. The only piece of information available about the objective \(F(z)\) is that it is Schur-convex on the feasible set
\[
Z = \{ z \in \mathbb{R}^N : z + m \preceq y \}. \tag{5.4}
\]
The definition of Schur-convexity suggests that posing the optimization problem (5.3) for a Schur-convex \(F(z)\) is intimately connected with a question of more fundamental nature: Does there exist a minimum element \(z^*\) in the set \(Z\) with respect to the partial order of majorization? In other words, is there a \(z^* \in Z\) such that \(z^* \preceq z\) for any \(z \in Z\)?

We shall seek the answer to this question by addressing the optimization problem (5.3) within the framework of convex optimization. In order to gain access to the tools of convex optimization, we shall initially compromise generality by constraining \(F(z)\) to the sub-class of Schur-convex functions that are symmetric, convex, and continuously differentiable on \(\mathbb{R}^N\).

To ease the following discussion, we shall introduce some notation. Given a vector \(a = (a_1, \ldots, a_N)\), denote the cumulative-sum vector as \(\tilde{a} = (\tilde{a}_1, \ldots, \tilde{a}_N)\) defined by \(\tilde{a}_k = \sum_{n=1}^{k} a_n\). For later notational simplicity we define \(\tilde{a}_0 = 0\). Denote by \(D_\downarrow\) the set of decreasing vectors,

\[\text{A function } F \text{ is symmetric if } F(z) = F(\Pi z) \text{ for any } z \in \mathbb{R}^N \text{ and any } N\text{-dimensional permutation matrix } \Pi. \]

\[\text{F is convex if } F(\theta z_1 + (1 - \theta)z_2) \leq \theta F(z_1) + (1 - \theta) F(z_2) \text{ for any two } z_1, z_2 \in \mathbb{R}^N \text{ and all } \theta \in [0, 1]. \]

\[\text{That all symmetric and convex functions are Schur-convex is shown in [MOA11, Prop. 3.2].}\]
\[ D_\downarrow = \{ \mathbf{a} \in \mathbb{R}^N : a_1 \geq ... \geq a_N \}, \] and the set of increasing vectors by \[ D_\uparrow = \{ \mathbf{a} \in \mathbb{R}^N : a_1 \leq ... \leq a_N \}. \]

Without loss of generality we assume that both \( \mathbf{y} \) and \( \mathbf{m} \) in (5.3) are decreasing vectors. Clearly, assuming \( \mathbf{y} \in D_\downarrow \) is not a restriction, since the order of majorization is unaffected by permutations. For \( \mathbf{m} \), the symmetry property of \( \mathcal{F}(\mathbf{z}) \) implies that problem (5.3) with an unordered \( \mathbf{m} \) can be solved via the equivalent problem using the decreasing rearrangement \( \mathbf{m}_\downarrow \) in place of \( \mathbf{m} \). Problem (5.3) also belongs to the class of convex optimization problems in the sense that the function \( \mathcal{F}(\mathbf{z}) \) to be minimized is convex, and the domain \( \mathcal{Z} \) is a convex set. The latter can be verified by equivalently writing the majorization constraint in (5.4) in terms of an affine equality constraint

\[ \tilde{z}_N + \tilde{m}_N = \tilde{y}_N, \quad (5.5) \]

together with a number of affine inequality constraints

\[ \sum_{n \in S} z_n + m_n \leq \tilde{y}_{|S|} \quad \forall \ S \subset \{1, \ldots, N\}, \quad (5.6) \]

in which \(|S|\) is the cardinality of \( S \).

### 5.1.1 Optimality Conditions

We shall derive optimality conditions that are sufficient: Any \( \mathbf{z}^\ast \) satisfying these optimality conditions is a global minimizer of (5.3) for any symmetric, convex, and continuously differentiable objective. To this end, we shall consider a relaxation of the same problem by extending the feasible region. We let \( \mathcal{Z}_{relax} \) exclude all inequality constraints in (5.6) except those corresponding to \( S = \{1, 2, \ldots, k\} \) for \( k = 1, \ldots, N - 1 \). Introducing the vector \( \mathbf{b} = \mathbf{y} - \mathbf{m} \), \( \mathcal{Z}_{relax} \supset \mathcal{Z} \) is defined as

\[ \mathcal{Z}_{relax} = \{ \mathbf{z} \in \mathbb{R}^N : \tilde{z} \leq \tilde{b}, \ \tilde{z}_N = \tilde{b}_N \}, \quad (5.7) \]

where the inequality \( \tilde{z} \leq \tilde{b} \) is taken componentwise. Note that if a minimizer \( \mathbf{z}^\ast \) of the relaxed problem

\[
\begin{align*}
\text{minimize} & \quad \mathcal{F}(\mathbf{z}) \\
\text{subject to} & \quad \tilde{z} \leq \tilde{b} \\
& \quad \tilde{z}_N = \tilde{b}_N
\end{align*}
\]

satisfies \( \mathbf{z}^\ast \in \mathcal{Z} \), then \( \mathbf{z}^\ast \) is a minimizer of the original problem (5.3) as well.

We begin our analysis of the relaxed problem with the Karush-Kuhn-Tucker (KKT) optimality conditions. The Lagrangian function of the relaxed problem is

\[ \mathcal{L} = \mathcal{F}(\mathbf{z}) + \lambda^T (\tilde{z} - \tilde{b}), \quad (5.9) \]
with Lagrangian multipliers $\lambda = [\lambda_1, \ldots, \lambda_N]^T$, which gives the KKT optimality conditions [BV04, Sec. 5.5],

$$\nabla F(z)_n + \sum_{i=n}^{N} \lambda_i = 0, \quad 1 \leq n \leq N \tag{5.10}$$

$$\lambda_n \geq 0, \quad 1 \leq n \leq N - 1 \tag{5.11}$$

$$\tilde{z}_n \leq \tilde{b}_n, \quad 1 \leq n \leq N - 1 \tag{5.12}$$

$$\tilde{z}_N = \tilde{b}_N \tag{5.13}$$

$$\lambda^T (\tilde{b} - \tilde{z}) = 0. \tag{5.14}$$

Since the problem is convex and all constraints are affine, the existence of a feasible point ensures that strong duality holds and that the KKT conditions are both sufficient and necessary for optimality [BV04, Sec. 5.2.3]. Since the feasible region is nonempty (for example, $z = b$ is feasible), we can find a minimizer of the relaxed problem by solving the KKT conditions. The KKT condition (5.10) above depends on the objective function $F(z)$ that is used. Next, we show in Theorem 5.1 the entire class of Schur-convex objectives can be treated in a unified way. This relies on the following auxiliary result, which is an immediate consequence of Schur’s condition, and stated here without proof.

**Lemma 5.1 (Schur’s condition, [MOA11, Thm. 3.A.4]).** If $F(z)$ is symmetric, convex, and continuously differentiable on $\mathbb{R}^N$, then for any $z \in \mathbb{R}^N$, $\nabla F(z)$ and $z$ are similarly ordered in the sense that $z_i \leq z_j$ implies $\nabla F(z)_i \leq \nabla F(z)_j$.

**Theorem 5.1.** Provided that $F(z)$ is symmetric, convex, and continuously differentiable, the KKT conditions of the relaxed problem are fulfilled for any $z$ satisfying the following conditions

**C-1.** $z \in D_\uparrow$,

**C-2.** $\tilde{z}_n \leq \tilde{b}_n, \quad 1 \leq n \leq N - 1$

**C-3.** $\tilde{z}_N = \tilde{b}_N$,

**C-4.** $z_n = z_{n+1}$ if $\tilde{z}_n < \tilde{b}_n, \quad 1 \leq n \leq N - 1$.

with associated multipliers

$$\lambda_k = \begin{cases} 
\nabla F(z)_{k+1} - \nabla F(z)_k, & \text{if } k = 1, \ldots, N - 1, \\
-\nabla F(z)_N, & \text{if } k = N.
\end{cases} \tag{5.15}$$
5.1. SKEWED MAJORIZATION CONSTRAINTS

Proof. Conditions C-2 and C-3 are identical to the corresponding KKT conditions (5.12) and (5.13), respectively, and (5.10) is satisfied by defining \( \lambda \) as in (5.15). By Schur’s condition, C-1 implies that

\[
\mathbf{z} \in \mathcal{D}^\uparrow \implies \nabla F(\mathbf{z}) \in \mathcal{D}^\uparrow \implies \lambda_k \geq 0,
\]

for all \( k = 1, \ldots, N-1 \), satisfying (5.11). If \( \tilde{z}_k < \tilde{b}_k \) for any \( k = 1, \ldots, N-1 \), then condition C-4 implies (5.14) via Schur’s condition as

\[
z_k = z_{k+1} \implies \nabla F(\mathbf{z})_k = \nabla F(\mathbf{z})_{k+1} \implies \lambda_k = 0.
\]

The proposed optimality conditions in Theorem 5.1 are hence sufficient to fulfill the KKT conditions. Since these are sufficient, any vector \( \mathbf{z}^* \) that satisfies conditions C-1 to C-4 must be a minimizer of the relaxed problem (5.8). The question now is whether such a point is also a minimizer of the original problem (5.3). Recall that this is true if \( \mathbf{z}^* \) satisfies \( \mathbf{z}^* + \mathbf{m} \preceq \mathbf{y} \), which holds if \( \mathbf{z}^* + \mathbf{m} \in \mathcal{D} \) since \( \tilde{\mathbf{z}}^* + \tilde{\mathbf{m}} \leq \tilde{\mathbf{y}} \). The following theorem reveals that the optimal \( \mathbf{z}^* + \mathbf{m} \) is in fact a decreasing vector, provided that \( \mathbf{m} \) and \( \mathbf{y} \) are ordered in decreasing order.

Theorem 5.2. If \( \mathbf{z}^* \) satisfies the optimality conditions in Theorem 5.1 for \( \mathbf{b} = \mathbf{y} - \mathbf{m} \) (with \( \mathbf{y}, \mathbf{m} \in \mathcal{D}^\downarrow \)), then \( \mathbf{z}^* + \mathbf{m} \in \mathcal{D}^\downarrow \).

Proof. Since the optimality conditions are satisfied, \( \tilde{z}_n^* + \tilde{m}_n \leq \tilde{y}_n \) for all \( 1 \leq n \leq N-1 \). Assume first that \( \tilde{z}_n^* + \tilde{m}_n < \tilde{y}_n \). Then

\[
z_n^* + m_n = z_{n+1}^* + m_n \geq z_{n+1}^* + m_{n+1}. \tag{5.17}
\]

If, on the other hand, \( \tilde{z}_n^* + \tilde{m}_n = \tilde{y}_n \), then

\[
\begin{align*}
\tilde{z}_{n-1}^* + \tilde{m}_{n-1} & \leq \tilde{y}_{n-1} \\
\tilde{z}_n^* + \tilde{m}_n & = \tilde{y}_n \\
\tilde{z}_{n+1}^* + \tilde{m}_{n+1} & \leq \tilde{y}_{n+1}
\end{align*}
\implies
\begin{align*}
z_n^* + m_n & \geq y_n \\
z_{n+1}^* + m_{n+1} & \leq y_{n+1}
\end{align*}
\implies
\begin{align*}
z_n^* + m_n & \geq z_{n+1}^* + m_{n+1}
\end{align*}
\]

where the last inequality uses the fact that \( y_n \geq y_{n+1} \). \( \square \)
5.1.2 Extended Applicability of the Optimality Conditions

Summarizing the findings up to this point, the optimality conditions in Theorem 5.1 are sufficient to find a minimizer of the optimization problem

\[
\begin{align*}
\text{minimize} & \quad F(z) \\
\text{subject to} & \quad z + m \preceq y,
\end{align*}
\]

provided that \( F(z) \) is symmetric, convex, and continuously differentiable. While these properties were necessary to find the optimality conditions within the framework of convex optimization, they are in fact overly restrictive. Next we show that the only required property of \( F(z) \) is that it is Schur-convex on the feasible set.

**Theorem 5.3.** If \( z^* \) satisfies the optimality conditions in Theorem 5.1 for \( b = y - m \) (with \( y, m \in D_\downarrow \)), then \( z^* \preceq z \) for any \( z \in \mathbb{R}^N \) satisfying \( z + m \preceq y \).

**Proof.** A proof is located in Appendix 5.B.1.

Interestingly, there is a companion result to Theorem 5.3 for optimizing a Schur-convex objective \( F(z) \) subject to a right-skewed majorization constraint:

\[
\begin{align*}
\text{minimize} & \quad F(z) \\
\text{subject to} & \quad m \preceq y - z.
\end{align*}
\]

In contrast to the previously studied optimization problem (5.18), the right-skewed majorization constraint corresponds to a *non-convex* feasible region. Nevertheless, the optimality conditions in Theorem 5.1 are sufficient for optimality for this optimization problem as well. Below we state this result without a proof, since it is almost identical to that of Theorem 5.3.

**Theorem 5.4.** If \( z^* \) satisfies the optimality conditions in Theorem 5.1 for \( b = y - m \) (with \( y, m \in D_\downarrow \)), then \( z^* \preceq z \) for any \( z \in \mathbb{R}^N \) satisfying \( m \preceq y - z \).

5.1.3 Algorithms producing the optimal solution

In this section we prove that the problem of finding a \( z \) 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 \( \mathbb{R}^2 \).
5.1. SKEWED MAJORIZATION CONSTRAINTS

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, Fig. 5.1 illustrates the distinction between a sequence of points, a simple polygon, and the convex hull under the simple polygon respectively.

**Theorem 5.5.** A $z$ satisfying the optimality conditions in Theorem 5.1 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 $\tilde{b}$ into $\mathbb{R}^2$ according to

$$p_n = (n, \tilde{b}_n), \quad n = 0, \ldots, N,$$

(5.20)

it is obvious that either $p_0 \ldots p_N$ is a simple polygon, or there are points above the line $p_0p_N$ that may be removed to form a simple polygon.

A convex-hull algorithm produces a simple, convex polygon $p_{n_0} \ldots p_{n_q}$, 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, or

$$\frac{\tilde{b}_{n_{r+1}} - \tilde{b}_{n_r}}{n_{r+1} - n_r} > \frac{\tilde{b}_{n_r} - \tilde{b}_{n_{r-1}}}{n_r - n_{r-1}}, \quad r = 1, \ldots, q - 1,$$

(5.21)
and the hull property that the straight line between neighboring points in the hull forms a lower bound:

\[
\tilde{b}_n \geq \tilde{b}_{n_r} + (n - n_r) \frac{\tilde{b}_{n_{r+1}} - \tilde{b}_{n_r}}{n_{r+1} - n_r}, \quad \begin{cases} r = 0, \ldots, q - 1, \\ n_r < n < n_{r+1}. \end{cases}
\] (5.22)

We are now ready to use the convex simple polygon, specified by the set of polygon indices \( I_A = \{n_0, \ldots, n_q\} \), to construct a vector \( z \) that satisfies the optimality conditions. First, we enforce equality at the points of the polygon, i.e., \( \tilde{z}_n = \tilde{b}_n \) if \( n \in I_A \). Then, in order to satisfy condition C-4 we make the slope of \( \tilde{z}_n \) constant for indices between polygon points,

\[
z_n = \frac{\tilde{b}_{n_{r+1}} - \tilde{b}_{n_r}}{n_{r+1} - n_r}, \quad \text{if} \quad n_r < n \leq n_{r+1}. \] (5.23)

Clearly, the vector \( z \) 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 \( N \in I_A \). Pick an arbitrary index, \( n \in \{1, \ldots, N - 1\} \), and identify the corresponding polygon point, \( r \), satisfying \( n_r < n \leq n_{r+1} \). If \( n \in I_A \) then \( \tilde{z}_n = \tilde{b}_n \), while if \( n \not\in I_A \) then by (5.22)

\[
\tilde{z}_n = \tilde{b}_n + (n - n_r) \frac{\tilde{b}_{n_{r+1}} - \tilde{b}_{n_r}}{n_{r+1} - n_r} \leq \tilde{b}_n.
\] (5.24)

Hence, C-2 is satisfied. Moreover, if \( n \not\in I_A \) then \( z_n = z_{n+1} \), while if \( n \in I_A \) then

\[
z_n = \frac{\tilde{b}_{n_r} - \tilde{b}_{n_r-1}}{n_r - n_{r-1}} < \frac{\tilde{b}_{n_{r+1}} - \tilde{b}_{n_r}}{n_{r+1} - n_r} = z_{n+1},
\] (5.25)

and we conclude that \( z \in D_\uparrow \) (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 5.A, tailored for solving the particular problem considered herein.

### 5.2 Optimization with Submodular Constraints

In this section, we shall briefly outline and explore an interesting connection between optimization with majorization constraints and a more
general class of optimization problems: optimization with submodular constraints. Such problems have been thoroughly investigated in the literature from an algorithmic point of view for the particular case of a separable objective function. The connection between majorization and submodular constraints is quite fruitful, since the algorithms can be made very efficient when particularized to majorization constraints due to the special symmetry of the majorization relation. To the author’s best knowledge, this approach is novel in the context of MIMO transceiver design.

The starting point is an optimization problem with a majorization constraint in its most simple form:

$$\min_{x} \mathcal{H}(x)$$

subject to  \( x \preceq y \)

In the previous section, we investigated in detail the case where \( \mathcal{H}(x) \) could be written as

$$\mathcal{H}(x) = \mathcal{F}(x - m)$$

in terms of a Schur-convex function \( \mathcal{F}(z) \) and a given vector \( m \). In this section, we shall instead focus on the case where \( \mathcal{H}(x) \) is a general separable function on the form

$$\mathcal{H}(x) = \sum_{n=1}^{N} h_n(x_n),$$

where each \( h_n(x) \) is assumed to be strictly convex and decreasing. We shall approach the optimization problem (5.26) by representing the majorization relation by a submodular base polyhedron. In the following definition, \( N = \{1, \ldots, N\} \) and \( 2^N \) denotes the set of all subsets of \( N \).

**Definition 5.2.** A set function \( \rho : 2^N \to \mathbb{R} \) satisfying the submodular property

$$\rho(S_1) + \rho(S_2) \geq \rho(S_1 \cup S_2) + \rho(S_1 \cap S_2),$$

for any two \( S_1, S_2 \in 2^N \), has a corresponding submodular base polyhedron \( B(\rho) \subset \mathbb{R}^N \) defined as

$$x \in B(\rho) \iff \begin{cases} \sum_{n \in S} x_n \leq \rho(S), & \forall S \in 2^N, \\ \sum_{n \in N} x_n = \rho(N). \end{cases}$$

In conventional terminology, Definition 5.2 refers to a submodular system \((\rho, 2^N)\) on \( N \).\(^2\) The connection between submodular base polyhedra

\(^2\)More general submodular systems \((\rho, \mathcal{L})\) can be similarly defined for a distributive lattice \( \mathcal{L} \subset 2^N \) by replacing \( 2^N \) by \( \mathcal{L} \) in Definition 5.2 (see [Fuj05] for further details).
and majorization was observed in [Fuj05, Sec. 2.3] and is presented below as a lemma.

**Lemma 5.2.** For a given \( y \in \mathbb{R}^N \), the set function

\[
\rho_y(S) = \sum_{n=1}^{\lfloor |S| \rfloor} y_{\downarrow n}, \quad S \in 2^N
\]

defines a submodular base polyhedron with the property that

\[
x \in B(\rho_y) \iff x \preceq y.
\]

The problem we set out to solve in (5.26) is hence equivalent to the following separable program with submodular constraints:

\[
\text{minimize} \quad \sum_{n=1}^{N} h_n(x_n), \\
\text{subject to} \quad x \in B(\rho_y).
\]

The problem we set out to solve in (5.26) is hence equivalent to the following separable program with submodular constraints:

\[
\text{minimize} \quad \sum_{n=1}^{N} h_n(x_n), \\
\text{subject to} \quad x \in B(\rho_y).
\]

A detailed exposition on optimization problems with general submodular constraints can be found in [Fuj05]. The algorithmic framework is based on two important contributions by Fujishige [Fuj80] and Groenevelt [Gro91], and we shall only focus on the latter here. Groenevelt proposed a decomposition algorithm to solve a convex, separable optimization problem over a polymatroid – a concept closely related to the submodular base polyhedron as defined in Definition 5.2. The algorithm was later extended to cover general submodular systems in [Fuj05].

The decomposition algorithm in [Gro91] was shown to be particularly suitable for generalized symmetric set functions \( \rho(S) \). One example of such a set function is when the value of \( \rho(S) \) only depends on \( |S| \): the number of elements in the set \( S \). Referring to Lemma 5.2, this is precisely the case for the set function representing the majorization constraint. In Algorithm 1 we have particularized the decomposition algorithm in [Gro91] to the set function in Lemma 5.2. For convenience, it is stated directly in terms of the vector \( y \). Allowing for recursive calls of the algorithm necessitates generalizing the problem formulation slightly. Below we define the following optimization problem over \( \{x_n\}_{n \in S} \) for \( S \subset \mathcal{N} \), which we shall refer to as \( P(S, \hat{y}_{\downarrow}) \) for a decreasing vector \( \hat{y}_{\downarrow} \in \mathbb{R}^{\lfloor |S| \rfloor} \).

\[
P(S, \hat{y}_{\downarrow}) : \quad \text{minimize} \quad \sum_{n \in S} h_n(x_n), \\
\text{subject to} \quad \{x_n\}_{n \in S} \preceq \hat{y}_{\downarrow}.
\]

The original problem in (5.26) corresponds to \( P(N, y_{\downarrow}) \).
5.2. OPTIMIZATION WITH SUBMODULAR CONSTRAINTS

Algorithm 1: A Decomposition Algorithm.

**Input**: Index set $S \subset \{1, \ldots, N\}$ and $\tilde{y}_1 \in \mathbb{R}^{\left|S\right|}$.

**Output**: A maximizer $\{x_n\}_{n \in S}$ of $P(S, \tilde{y}_1)$.

1. Find a solution $\{z_n^*\}_{n \in S}$ to the single-constraint problem

   $$\begin{align*}
   &\text{minimize} \quad \sum_{n \in S} h_n(z_n), \\
   &\text{subject to} \quad \sum_{n \in S} z_n = \sum_{n=1}^{\left|S\right|} \tilde{y}_n
   \end{align*}$$

2. Order $S = \{s_1, \ldots, s_{\left|S\right|}\}$ such that $z_{s_1}^* \geq \ldots \geq z_{s_{\left|S\right|}}^*$

3. for $i = 1$ to $\left|S\right|$ do

   4. \hspace{1em} $v_{s_i} = \min \left( z_{s_i}^*, \sum_{j=1}^{i} \tilde{y}_j - \sum_{j=1}^{i-1} v_{s_j} \right)$

5. Find maximum $m$ such that $\sum_{i=1}^{m} \tilde{y}_i = \sum_{i=1}^{m} v_{s_i}$

6. if $m = \left|S\right|$ then

   7. \hspace{1em} $x_n^* = z_n^*, \quad \forall n \in S$

8. else

   9. Define $S^1 = \{s_1, \ldots, s_m\}$ and $\tilde{y}_1^1 = [\tilde{y}_{s_1}, \ldots, \tilde{y}_{s_m}]^T$. Find the solution $\{x_n^*\}_{n \in S^1}$ of $P(S^1, \tilde{y}_1^1)$

10. Define $S^2 = \{s_{m+1}, \ldots, s_{\left|S\right|}\}$ and $\tilde{y}_1^2 = [\tilde{y}_{s_{m+1}}, \ldots, \tilde{y}_{s_{\left|S\right|}}]^T$. Find the solution $\{x_n^*\}_{n \in S^2}$ of $P(S^2, \tilde{y}_1^2)$

Solving problem (5.26) using the decomposition algorithm starts with disregarding all hidden linear constraints in $x \preceq y$ except for $\sum_{n=1}^{N} x_n = \sum_{n=1}^{N} y_n$. If the solution to the single-constraint problem is feasible for the original problem as well (this is checked on line 6), the minimizer has been found and the algorithm stops. Otherwise, the procedure on lines 2–5 has cleverly identified two subproblems that will lead to the desired solution (see [Gro91] for further details).

In total, at most $N - 1$ non-trivial executions of the algorithm are necessary to solve the problem. For each instance, the ordering procedure on line 2 is an $\mathcal{O} (\left|S\right| \log \left|S\right|)$ operation, and the complexity of solving (5.26) using the decomposition algorithm is bounded by $\mathcal{O} (N^2 \log N)$ [Gro91]. The simple single-constraint problem on line 1 can in general be solved using a bisection search. If each $h_n (x)$ is differentiable and strictly convex the solution for $S = N$ is implicitly given by

$$h'_1 (z_1) = \ldots = h'_N (z_N) = \lambda^*,$$
where $\lambda^*$ is chosen to satisfy $z_n = \sum_{n=1}^{N} y_{\downarrow n}$.

5.3 Statistical Precoding for ZF-DF Equalization

In this section we describe several applications of the theoretical results derived in the preceding sections. We shall return to the framework for statistical precoding for ZF-DF equalization in Chapters 3 and 4, assuming that the transmitter has access to long-term, statistical CSI while the receiver has perfect, short-term CSI. With $n_T$ transmit antennas, $n_R$ receive antennas, and $n_L \leq \min(n_T, n_R)$ data streams for spatial multiplexing, system performance depends on the subchannel SNR vector $\gamma = (\gamma_1, \ldots, \gamma_{n_L})$, which has been investigated under the following two statistical channel models for the MIMO channel $H \in \mathbb{C}^{n_R \times n_T}$:

- **Separable-Correlation Rayleigh Model:** $H = R_{1/2} Z R_{1/2}^T$,

- **Double-Scattering Model:** $H = R_{1/2} Z_R Z_R^* R_{1/2}^T$,

where the matrices $Z \in \mathbb{C}^{n_R \times n_T}$, $Z_R \in \mathbb{C}^{n_R \times n_S}$, and $Z_S \in \mathbb{C}^{n_S \times n_T}$ all have i.i.d. $\mathcal{CN}(0, 1)$ entries. For both channel models, we showed in Theorem 4.1 that the SNR vector can be factored as $\gamma \sim \bar{\gamma} \odot x$,

in terms of a random component, the normalized SNR $\bar{\gamma}$, and a deterministic component, the power allocation vector $x = l^2(P^H R_T P)$. An optimization problem for the precoder $P \in \mathbb{C}^{n_T \times n_L}$, designed to maximize system performance subject to a total transmit-power budget $P_{\text{max}}$ was posed in Design Formulation 3.1 for a general cost function $F_\gamma(x)$ as

$$\minimize_{P} \quad F_\gamma(x)$$

subject to

$$x = l^2(P^H R_T P),$$

$$\text{tr} \left( P^H P \right) \leq P_{\text{max}}.$$ (5.28)

Aided by the generalized triangular decomposition [JHL08], the optimization problem was reformulated in Design Formulation 3.2. With $\bar{x} = \log(x)$ and $\bar{p} = \log(p)$, where $p \in \mathbb{R}^{n_L}$ denotes the eigenvalues of $P^H P$, the transformed optimization problem was given by

$$\minimize_{\bar{x}, \bar{p}} \quad F_\gamma(e^{\bar{x}})$$

subject to

$$\bar{x} \preceq \bar{p} + \lambda^*_T,$$

$$\sum_{k=1}^{n_L} e^{\bar{p}_k} \leq P_{\text{max}}.$$ (5.28)
Here, $\lambda^l_T = (\log(\lambda_{T,1}), \ldots, \log(\lambda_{T,n_L}))$, and $\lambda_{T,1} \geq \ldots \geq \lambda_{T,n_L}$ are the $n_L$ largest eigenvalues of $R_T$.

### 5.3.1 MSE-Based Cost Functions

Here we shall take a closer look at the general class of MSE-based class of cost functions as described in detail in Sections 3.2.2 and 3.3.2. With the inverse relation between the MSE and the SNR for ZF receivers, $f_{\text{MSE}}(\gamma_k) = \gamma_k^{-1}$, the subchannel MSEs can be combined as

$$F_{\gamma} = \left\| \left\{ E\left[ \alpha_k f_{\text{MSE}}(\gamma_k) \right] \right\}_k \right\|^q_{n_L},$$

where $||z||_q = (\sum_k |z_k|^q)^{1/q}$. Assuming nonnegative weights, this class of cost functions includes the following two examples:

$$F_{\gamma} = \begin{cases} \sum_k E\left[ \alpha_k f_{\text{MSE}}(\gamma_k) \right], & \text{if } q = 1, \\ \max_k \left\{ E\left[ \alpha_k f_{\text{MSE}}(\gamma_k) \right] \right\}, & \text{if } q \to \infty. \end{cases}$$

By raising the cost function in (5.28) to the power of $q$, the optimization problem becomes

$$\begin{aligned} \text{minimize} & \quad \sum_{k=1}^{n_L} e^{-q(\tilde{x}_k - \tilde{\alpha}_k)} \\ \text{subject to} & \quad \tilde{x} \preceq \tilde{p} + \lambda^l_T, \\ & \quad \sum_{k=1}^{n_L} e^{p_k} \leq P_{\max}. \end{aligned}$$

Here, $\tilde{\alpha}_k = \log(\alpha_k E[\tilde{\gamma}_k^{-1}])$. In case of the transmit-correlated Rayleigh model $E[\tilde{\gamma}_k^{-1}] = (n_R - n_L + k - 1)^{-1}$ for $n_L < n_R$. For the more generally correlated Rayleigh or double-scattering channel models, $E[\tilde{\gamma}_k^{-1}]$ can be well approximated using the techniques developed in Chapter 4.

Interestingly, there is an almost closed-form solution to the optimization problem (5.30). This solution can be expressed using the following notion of a convex-hull solution as defined next.

**Definition 5.3.** For a given vector $y \in \mathbb{R}^N$, the convex-hull solution $\xi = \xi(y) \in \mathbb{R}^N$ is an increasing vector that satisfies $\sum_{k=1}^N \xi_k - y_k = 0$, and for any $n = 1, \ldots, N - 1$, $\sum_{k=1}^n \xi_k - y_k \leq 0$ and

$$\sum_{k=1}^n \xi_k - y_k < 0 \implies \xi_n = \xi_{n+1}.$$
This definition relies on the key result in Theorem 5.5. We may use any algorithm that identifies the convex hull of a simple polygon in \( \mathbb{R}^2 \) to extract \( \xi \) from \( y \), e.g. [MA79]. We advocate using the algorithm supplied in Appendix 5.A, which is tailored to extracting convex-hull solutions. This algorithm is efficient due to its \( \mathcal{O}(N) \) complexity, and is readily implemented using only a few lines of code.

Transforming the optimization problem (5.30) by a change of variables \( \bar{z} = \bar{x} - \bar{\alpha} \) leads to optimization over the doubly skewed majorization constraint \( \bar{z} + \bar{\alpha} \preceq \bar{p} + \lambda_T^l \). The objective is Schur-convex in \( \bar{z} \), and the power constraint is Schur-convex in \( \bar{p} \). It follows by Theorems 5.3 and 5.4 that there is a minimizing pair \((\bar{x}^*, \bar{p}^*)\) satisfying both \( \bar{x}^* - \bar{\alpha} = \xi(\bar{p}^* + \lambda_T^l - \bar{\alpha}) \) and \( \bar{p}^* = -\xi(\lambda_T^l - \bar{x}^*) \), provided that \( \bar{\alpha} \) is decreasing.

Interestingly, the minimizing pair can also be expressed in terms of a single convex-hull solution \( \xi(\lambda_T^l - \bar{\alpha}_\downarrow) \), by exploiting the specific cost function and power constraint in the optimization problem. This convex-hull solution can be interpreted either as the minimizing \( \bar{x} - \bar{\alpha} \) given a uniform power allocation \( (\bar{p} = 0) \), or as the most power-efficient \( -\bar{p} \) given equal subchannel costs \( \bar{x} = \bar{\alpha} \). The minimizing pair \((\bar{x}^*, \bar{p}^*)\) of (5.30) is expressed in the following theorem.

**Theorem 5.6.** Let \( \xi \) be the convex-hull solution of \( \lambda_T^l - \bar{\alpha}_\downarrow \). Then the minimizer of (5.30) is given by

\[
\bar{x}_{\downarrow k}^* = \bar{\alpha}_\downarrow k + \frac{1}{q + 1} \xi_k + \delta, \\
\bar{p}_k^* = -\frac{q}{q + 1} \xi_k + \delta, 
\]

where \( \delta = \log \left( P_{\text{max}} / \sum_{\nu=1}^{n_L} e^{-\frac{q}{q+1} \xi_\nu} \right) \), and \( \bar{x}^* \) should be ordered as \( \bar{\alpha} \).

**Proof.** Appendix 5.B.2 contains the proof, which identifies the KKT conditions, constructs optimal dual variables and verifies that every condition is met. \( \square \)

The result in Theorem 5.6 enables solving the power allocation problem efficiently by computing a convex-hull solution using, for example, the algorithm in Appendix 5.A; the exact optimum is easily extracted with \( \mathcal{O}(n_L \log(n_L)) \) complexity, including sorting \( \bar{\alpha} \) into \( \bar{\alpha}_\downarrow \). Alternatively, as mentioned in Chapter 3, the MSE-based design problem can also be solved

\(^3\) Its input is the cumulative-sum vector \( \tilde{y} = [\tilde{y}_0 \ldots \tilde{y}_N]^T \) (corresponding to \( x \) in Algorithm 5.1). The algorithm returns a number of indices \( i[0], \ldots, i[n_L] \), where \( 0 = i[0] < \ldots < i[n_L] = N \) and \( 1 \leq n_L \leq N \), for which \( \tilde{\xi}_{\nu[i]} = \tilde{y}_{\nu[i]} \). The \( n \)th component \( \xi_n \) of the convex-hull solution \( \xi \) is extracted as \( \xi_n = (\tilde{y}_{i[r+1]} - \tilde{y}_{i[r]})/(i[r+1] - i[r]) \), where for each \( n \), \( r \) is the unique integer satisfying \( i[r] < n \leq i[r+1] \).
using Design Formulation 3.5 by implementing an interior-point algorithm. Such an algorithm is more involved to implement, and iterates towards a minimizer. In each iteration a Newton step needs to be calculated, which is typically an $\mathcal{O}(n^3)$ operation. This approach was employed in [LZW09], which studied the particular MSE-based design corresponding to a uniform weight vector $\alpha$, $q = 1$, and dimensions satisfying $n_R > n_T = n_L$. Our contribution in Theorem 5.6 implies solving not only the problem in [LZW09], but also an entire class of MSE-based design problems using the computationally efficient algorithm in Appendix 5.A, thereby reducing the complexity to $\mathcal{O}(n_L \log(n_L))$.

### 5.3.2 Unitary Precoding

Up to this point, precoders have been optimized subject to a constraint on the total transmission power. However, due to hardware constraints it might be necessary to also limit the maximum transmitted power per antenna. This can be ensured in a strict sense by considering unitary precoding. We recall that the transmitted signal $x \in \mathbb{C}^{n_T}$ is constructed using linear precoding of the data signal vector $s \in \mathbb{C}^{n_L}$ as

$$x = Ps.$$  

The transmitted power for each antenna is then given by the diagonal elements of the matrix

$$\mathbb{E}[xx^H] = PP^H,$$

assuming that $s$ is normalized as $\mathbb{E}[ss^H] = I$. With one data stream per transmit antenna $n_L = n_T$, we may assert that

$$PP^H = I\frac{P_{\text{max}}}{n_T},$$

which ensures that the total transmission power $P_{\text{max}}$ is evenly distributed among the antennas.$^4$

Unitary precoding can easily be enforced in the unconstrained MSE-minimizing optimization problem for statistical precoding in (5.30), assuming that $n_L = n_T < n_R$. Since $\bar{p}$ is the logarithm of the eigenvalues of $PP^H$, setting $\bar{p} = (1, \ldots, 1) \cdot \log(P_{\text{max}}/n_T)$ results in

$$\min_{\bar{x}, \bar{p}} \sum_{k=1}^{n_T} e^{-q(x_k - \bar{\alpha}_k)}$$

subject to $\bar{x} \preceq \bar{p} + \bar{\lambda}_T$.

---

$^4$Note that unitary precoding is overly restrictive for achieving a uniform power distribution among the antennas. It is only necessary to require that the rows of $P$ have the same norm. However, we restrict ourselves to the class of unitary precoders since it leads to a tractable solution.
With a change of variables $\mathbf{z} = \bar{\mathbf{x}} - \bar{\alpha}$, this refers to the minimization of a Schur-convex objective under a skewed majorization constraint, which was dealt with in Section 5.1. The minimizer is given by the convex-hull solution $\mathbf{x}^\ast - \bar{\alpha} = \xi (\bar{\mathbf{p}} + \bar{\lambda}^T - \bar{\alpha})$, assuming that $\bar{\alpha}$ is decreasing. Again we note that $\xi (\bar{\mathbf{p}} + \bar{\lambda}^T - \bar{\alpha})$ can be extracted in $O(n_T)$ time using the algorithm in Appendix 5.A.

The MSE-based objective $\left\| \left\{ \mathbb{E} \left[ \alpha_k f^{\text{MSE}} (\gamma_k) \right] \right\}_{k=1}^{n_T} \right\|_q$ represented by (5.32) can be exchanged for other performance measures of interest. If $h_k (\bar{x}_k)$ denotes an arbitrary cost function for subchannel $k$, the sum-cost minimizing optimization problem is given by

$$\begin{align*}
\text{minimize} & \quad \sum_{k=1}^{n_T} h_k (\bar{x}_k) \\
\text{subject to} & \quad \mathbf{x} \preceq \bar{\mathbf{p}} + \bar{\lambda}^T .
\end{align*}$$

(5.33)

As shown in Section 5.2, any such problem can be posed as a separable optimization problem with submodular constraints. Assuming that each $h_k (\bar{x}_k)$ is convex, a minimizer can be computed efficiently using the decomposition algorithm in Algorithm 1. Reviewing the list of example cost functions in Section 3.2.2, the separable form in (5.33) directly applies to the average bit-error probability. With slight reformulations, the joint symbol error probability as well as the outage formulation can also be represented by a separable, convex objective.

It is also possible to consider unitary precoding under the assumption that the transmitter has access to perfect short-term CSI. Interestingly, the corresponding optimization problems for both ZF-DF and MMSE-DF receivers with perfect transmitter CSI are very similar to the ones in (5.32) and (5.33) for the ZF-DF receiver and long-term, statistical CSI at the transmitter. The corresponding MSE-based design with perfect CSI has a convex-hull solution that can be readily extracted, and separable programs can be solved using the decomposition algorithm. For further details, we refer the interested reader to [BJOJ08].

### 5.4 Summary

In this chapter, we have investigated optimization problems involving majorization constraints in their own respect. When optimizing a Schur-convex function with respect to a skewed majorization constraint, it was shown that there is a common solution for the entire class of cost functions. It is also shown that the optimization problem is equivalent to identifying the convex hull under a simple polygon defined by the constraint parameters. We have also shown that a majorization constraint is a special case
5.4. SUMMARY

of more general submodular constraints. The case of minimizing a general separable and convex function subject to such a constraint has been studied in the literature, and a decomposition algorithm that solves the problem can be very efficiently stated when particularized to a majorization constraint.

The general optimization problems could be directly applied to a particular case of statistical precoder design: unitary precoding that forces the total transmit power to be evenly distributed over the antennas. The corresponding unconstrained optimization problem involved the optimization over a doubly skewed majorization constraint. Considering an MSE-based cost function, we showed that this problem was equivalent to that of minimizing a Schur-convex function over a single skewed majorization constraint, and the problem reduced to simply identifying the convex hull of a set of points in $\mathbb{R}^2$. It was observed that in the numerical evaluation in Section 3.4 that MSE-based precoders displayed only a marginal performance degradation to the optimal precoders. This observation makes MSE-based precoders highly attractive, since they can be efficiently determined without resorting to standard algorithms of convex optimization.
5.A A Convex-Hull Algorithm

Algorithm 5.1 determines the indices that correspond to points on the one-sided convex hull under \( x \). The vector of indices in the convex hull is denoted \( i \). After the algorithm has terminated it contains \( m + 1 \leq N \) elements. Line 1 initializes the vector with the first index 0 (remember, the first and the last points of \( x \) are always in the convex hull). In line 4, a loop over the elements of \( x \) is commenced. If the point at \( x_j \) is convex given that straight lines are drawn between \( x_{i[m]} \), \( x_j \), and \( x_N \), then the index \( j \) must be appended to the vector of hull-indices, \( i \). 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 \( i \) if they exist. Finally on line 14 the last point in \( x \) 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 \( O(N) \).

Algorithm 5.1

\begin{verbatim}
1:  i[0] ← 0
2:  N ← length(x) − 1
3:  m ← 0
4:  for j = 1 to N − 1 do
5:      if \( \frac{x_{i[m]} - x_j}{j-1[m]} > \frac{x_i - x_N}{N-j} \) then
6:          while m > 0 and \( \frac{x_{i[m-1]} - x_{i[m]}}{i[m]-i[m-1]} \leq \frac{x_j - x_{i[m]}}{j-1[m]} \) do
7:             m ← m − 1
8:          end while
9:      m ← m + 1
10:     i[m] ← j
11: end if
12: end for
13: m ← m + 1
14: i[m] ← N
\end{verbatim}
5.B A Collection of Proofs

5.B.1 Proof of Theorem 5.3

It suffices to compare \( z^* \) with increasing vectors \( z \) satisfying \( z + m \preceq y \). This follows by two observations. Firstly, for the increasing rearrangement \( z_\uparrow \) of \( z \), \( z^* \preceq z \) if and only if \( z^* \preceq z_\uparrow \). Secondly, if \( z + m \preceq y \) holds for \( z \), then \( z_\uparrow + m \preceq y \) holds for \( z_\uparrow \); this follows from the fact that \( z_\uparrow + m \preceq z + m \) for a decreasing \( m \) (see [MOA11, Prop. 6.A.2]).

We consider then an increasing \( z \) satisfying \( z + m \preceq y \), as well as an increasing \( z^* \) (by C-1). For increasing vectors, \( z^* \preceq z \) is equivalent to

\[
\tilde{z}_k^* \geq \tilde{z}_k, \quad k = 1, \ldots, N - 1, \tag{5.34}
\]

in conjunction with \( \tilde{z}_N^* = \tilde{z}_N \). The equality constraint holds by construction: \( \tilde{z}_N^* = \tilde{b}_N \) by C-3 and \( \tilde{z}_N = \tilde{y}_N - \tilde{m}_N = \tilde{b}_N \) by \( z + m \preceq y \). Next, we focus on the inequalities in (5.34). First we show that \( \tilde{z}_k^* \geq \tilde{z}_k \) for any \( k \) in the index set

\[
\mathcal{I} = \{ k \in \{0, \ldots, N\} : \tilde{z}_k^* = \tilde{b}_k \}.
\]

Irrespective of the order of \( z + m \), the majorization relation \( z + m \preceq y \) ensures that \( \tilde{z}_k + \tilde{m}_k \leq \tilde{y}_k \), or equivalently \( \tilde{z}_k \leq \tilde{b}_k \), for \( k = 1, \ldots, N - 1 \). For any \( k \in \mathcal{I} \), we then have

\[
\tilde{z}_k^* = \tilde{b}_k \geq \tilde{z}_k.
\]

Next, we show by contradiction that \( \tilde{z}_k^* \geq \tilde{z}_k \) holds for \( k \notin \mathcal{I} \) as well. Let \( n \) be the smallest index such that \( \tilde{z}_n^* < \tilde{z}_n \). Then there exist \( i, j \in \mathcal{I} \) such that \( i < n < j \), and \( \tilde{z}_l^* \leq \tilde{b}_l \) for all \( l \) such that \( i < l < j \). Since \( z \) is increasing and \( \tilde{z}_n^* < \tilde{z}_n \),

\[
\tilde{z}_j = \tilde{z}_n + (\tilde{z}_j - \tilde{z}_n) \\
\geq \tilde{z}_n + (j - n)z_n \\
> \tilde{z}_n^* + (j - n)z_n.
\]

Also, using \( \tilde{z}_{n-1}^* \geq \tilde{z}_{n-1} \) (by the definition of \( n \)) and \( \tilde{z}_n^* < \tilde{z}_n \) it follows that \( z_n^* < z_n \), and in turn,

\[
\tilde{z}_j > \tilde{z}_n^* + (j - n)z_n^* \\
= \tilde{z}_j^*,
\]

where the last equality is implied by C-4 stating that \( z_n^* = \ldots = z_j^* \). This contradicts our previous conclusion that \( \tilde{z}_j \leq \tilde{z}_j^* \) for \( j \in \mathcal{I} \). Hence, since there can be no \( n \) such that \( \tilde{z}_n^* < \tilde{z}_n \), (5.34) follows and the proof is concluded.
5.B.2 Proof of Theorem 5.6

We assume without loss of generality that $\bar{\alpha}$ is decreasing, and by the discussion in Section 3.3.1 we may consider the alternative formulation

$$\minimize_{\bar{x}, \bar{p}} \sum_{k=1}^{n_L} e^{-q(\bar{x}_k - \bar{\alpha}_k)}$$

subject to $\bar{x}_n \geq \bar{x}_{n+1}$, $n = 1, \ldots, n_L - 1,$

$$\sum_{k=1}^{n} \bar{x}_k \leq \sum_{k=1}^{n} \bar{p}_k + \bar{\lambda}_{T,k}, \quad n = 1, \ldots, n_L - 1,$$

$$\sum_{k=1}^{n_L} \bar{x}_k = \sum_{k=1}^{n_L} \bar{p}_k + \bar{\lambda}_{T,k},$$

$$\sum_{k=1}^{n_L} e^{\bar{p}_k} \leq P_{\text{max}}.$$

First, we check the feasibility of the proposed $\bar{x}^*$ and $\bar{p}^*$ in Theorem 5.6. Since $\xi$ is the convex-hull solution of $y = \bar{\lambda}_T - \bar{\alpha}$, and $\bar{x}_n - \bar{p}_n - \bar{\lambda}_{T,n} = \xi_n - (\bar{\lambda}_{T,n} - \bar{\alpha}_n)$, it is evident by Definition 5.3 that $\xi$ is tailored to meet the feasibility constraints

$$\sum_{n=1}^{k} \bar{x}_n - \bar{p}_n - \bar{\lambda}_{T,n} \leq 0, \quad k = 1, \ldots, n_L - 1, \quad (5.35)$$

and ensures equality for $k = n_L$. Also, we get $\sum_{k=1}^{n_L} e^{\bar{p}_n} = P_{\text{max}}$, and it remains to verify that $\bar{x}^*$ is decreasing. Note that $\bar{x}_k^* - \bar{x}_{k+1}^* = \bar{\alpha}_k - \bar{\alpha}_{k+1} + \frac{1}{q+1} (\xi_k - \xi_{k+1})$. Assume first that, for some $k$, $(5.35)$ holds with strict inequality. Then $\xi_k = \xi_{k+1}$ by Definition 5.3, and $\bar{x}_k^* - \bar{x}_{k+1}^* = \bar{\alpha}_k - \bar{\alpha}_{k+1} \geq 0$. If, instead, $(5.35)$ holds with equality, then since $\xi$ is increasing,

$$\bar{x}_k^* - \bar{x}_{k+1}^* \geq 2 \sum_{n=1}^{k} \xi_n + \bar{\alpha}_n - \sum_{n=1}^{k+1} \xi_n + \bar{\alpha}_n - \sum_{n=1}^{k-1} \xi_n + \bar{\alpha}_n$$

$$\geq 2 \sum_{n=1}^{k} \bar{\lambda}_{T,n} - \sum_{n=1}^{k+1} \bar{\lambda}_{T,n} - \sum_{n=1}^{k-1} \bar{\lambda}_{T,n} \geq \bar{\lambda}_{T,k} - \bar{\lambda}_{T,k+1} \geq 0.$$

We form the Lagrangian

$$\mathcal{L} = \sum_{k=1}^{n_L} \left( e^{-q(\bar{x}_k - \bar{\alpha}_k)} + \sum_{n=1}^{k} \mu_n (\bar{x}_n - \bar{p}_n - \bar{\lambda}_{T,n}) \right) + \nu \left( \sum_{k=1}^{n_L} e^{\bar{p}_k} - P_{\text{max}} \right).$$

Vanishing derivatives at the optimum gives the conditions $-qe^{-q(\bar{x}_k - \bar{\alpha}_k)} + \sum_{n=k}^{n_L} \mu_n = 0$ and $\nu e^{\bar{p}_k} - \sum_{n=k}^{n_L} \mu_n = 0$ for $k = 1, \ldots, n_L$. It is easy to
check that with $\bar{x}^*$ and $\bar{p}^*$ in Theorem 5.6, the optimal $\mu$ and $\nu$ are then consistently determined as

$$\mu_k = qe^{-q\delta} \left( e^{-\frac{q}{\bar{x}_k}} - e^{-\frac{q}{\bar{x}_{k+1}}} \right)$$

(5.36)

for $k = 1, \ldots, n_L - 1$, and $\mu_{n_L} = qe^{-q\left(\frac{1}{\bar{x}_{n_L}} + \delta\right)}$, and $\nu = qe^{-(q+1)\delta}$. These dual variables are feasible, since $\nu \geq 0$ and, due to $\xi$ being increasing, $\mu_k \geq 0$ for $k = 1, \ldots, n_L - 1$.

It remains to verify the complementary slackness $\mu_k \sum_{n=1}^{k} (\bar{x}_n^* - \bar{p}_n^* - \bar{\lambda}_{T,n}) = 0$. Since $\bar{x}_n^* - \bar{p}_n^* - \bar{\lambda}_{T,n} = \xi_n - (\bar{\lambda}_{T,n} - \bar{\alpha}_n)$, Definition 5.3 implies that either the sum is zero, and complementary slackness holds, or $\xi_k = \xi_{k+1}$. But the latter implies that $\mu_k = 0$ due to (5.36), which also implies complementary slackness.
Chapter 6

Precoding and Ordering in the Multi-User Uplink

This chapter generalizes and applies the methods in previous chapters to multi-user MIMO uplink transceiver design. In this context, the decoding order plays an important role, since the benefit of interference cancellation can be regarded as a resource to be allocated to different users. Since optimal precoders are dependent on the decoding order selected, these need to be jointly optimized.

6.1 System Model

The multi-user MIMO uplink, as illustrated in Figure 6.1, regards simultaneous transmission from \( K \) multi-antenna users to a base station (BS) equipped with \( n_R \) antennas. User \( k \) multiplexes a single data stream onto \( n_{L_k} \) substreams, which are independently encoded, and represented (at a given time instant) by the random signal vector \( s_k = [s_{k,1} \ldots s_{k,n_{L_k}}]^T \).

Figure 6.1: Model of a multi-user MIMO uplink, with decision feedback equalization at the BS.
satisfying $E[s_k s_k^H] = I$. A linear precoder $P_k \in \mathbb{C}^{n_T k \times n_L k}$ is applied on $s_k$ prior to transmission using $n_{T k} \geq n_{L k}$ antennas. The wireless MIMO link connecting user $k$ to the BS is embodied in the channel matrix $H_k \in \mathbb{C}^{n_R \times n_T k}$ under the assumptions of linear and narrowband communications. The received vector $y \in \mathbb{C}^{n_R}$ at the BS is given by the transmission model

$$y = \sum_{k=1}^{K} H_k P_k s_k + n$$

(6.1)

where $n \sim \mathcal{CN}(0, I)$ denotes both spatially and temporally white normal-distributed noise.

### 6.1.1 The Decision-Feedback Equalizer

Albeit geographically separated, the $K$ users can be seen as a single user from the receiver’s point of view. Collecting the $n_L = \sum_{k=1}^{K} n_{L k}$ data streams in a stacked signal vector $s = [s_1^T \ldots s_K^T]^T \in \mathbb{C}^{n_L}$, the virtual sole user utilizes all $n_T = \sum_{k=1}^{K} n_{T k}$ antennas via a block-diagonal precoder $P = \text{diag}(P_1, \ldots, P_K) \in \mathbb{C}^{n_T \times n_L}$ prior to transmission over the MIMO channel $H = [H_1 \ldots H_K] \in \mathbb{C}^{n_R \times n_T}$. In this notation the transmission model becomes $y = H P s + n$, which is the single-user case in (6.1).

The signal processing of the ZF-DF receiver was described in detail in Section 2.2 and Section 3.1, assuming that the data streams were enumerated according to their placements in the decoding order. Here we shall represent an arbitrary decoding order by an $n_L \times n_L$ permutation matrix $\Pi$, having entries in $\{0, 1\}$ and exactly one 1 in each row and each column. We instead assume that $\Pi s$ is decoded from the first component to the last. The ZF-DF receive matrices in Figure 6.1 can then be expressed in terms of the QL decomposition $Q L = H P \Pi^T$, where $Q \in \mathbb{C}^{n_R \times n_L}$ has orthonormal columns and $L \in \mathbb{C}^{n_L \times n_L}$ is lower triangular with positive diagonal elements:

$$W_{ZF} = Q D_L^{-1} \Pi, \quad B_{ZF} = \Pi^T D_L^{-1} L \Pi - I,$$

where $D_L$ is a diagonal matrix with the same diagonal as $L$. With $\bar{s} = W_{ZF}^H y - B_{ZF} \hat{s}$, and the conventional assumption on error-free feedback, the effective transmission model after successful interference nulling and cancelling is simply given by

$$\bar{s} = s + \tilde{n},$$

with a normal-distributed noise term $\tilde{n} = W_{ZF}^H n$ having diagonal covariance.
6.1. SYSTEM MODEL

The ZF-DF receiver decomposes the multi-user MIMO uplink into \( n_L \) non-interfering subchannels, which are parallel in the sense that the noise components are statistically independent. This point is illustrated to the left in Figure 6.2 in a two-user uplink with two subchannels per user, in which the decoding order is implicit; the assignment of subchannels to placements in the decoding order is illustrated to the right in Figure 6.2. Here, and in the following, \( \gamma_{k,n} \) denotes the SNR on the subchannel associated with \( s_{k,n} \). Collecting user \( k \)'s SNRs as \( \gamma_k = [\gamma_{k,1} \ldots \gamma_{k,n_L}]^T \), the system SNR vector \( \gamma = [\gamma_1^T \ldots \gamma_K^T]^T \) forms a basis for evaluating performance in the system.

**6.1.2 Channel Model and Information**

We focus on multipath-fading channels by adopting a statistical model for the users’ channel matrices. Channel correlations induced by insufficient antenna spacing at user \( k \) are represented by a positive semi-definite transmitter-side channel-correlation matrix \( R_{Tk} \in \mathbb{C}^{n_T k \times n_T k} \). The following two multi-user MIMO-channel models will be used.

- **General channel model:** The user MIMO channels \( H_1, \ldots, H_K \) are statistically independent and each can be factored as \( H_k = \tilde{H}_k R_{Tk}^{1/2} \), where \( \tilde{H}_k \) is statistically invariant to unitary transformations on the right-hand side.
• Transmit-correlated Rayleigh model: This instance of the general channel model assumes that $\tilde{H}_k$ is a standard complex normal-distributed matrix, i.e., $\text{vec}(\tilde{H}_k) \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$ for $k = 1, \ldots, K$.

The general channel model is in line with two common single-user MIMO channel models for the non-line-of-sight case, both discussed in detail in Chapter 2. The first is the Rayleigh-fading channel model with separable transmit and receive correlation, in which $H_k = R_{Rk}^{1/2} \tilde{H}_k R_{Tk}^{1/2}$ for a standard normal-distributed $\tilde{H}_k \in \mathbb{C}^{n_R \times n_T}$, and a positive semi-definite receiver-side channel-correlation matrix $R_{Rk} \in \mathbb{C}^{n_R \times n_R}$. The transmit-correlated Rayleigh model is an instance of the separable Rayleigh model with $R_{Rk} = \mathbf{I}$, and assumes well-separated elements in the BS antenna array. Another channel model that fits into the general definition is the double-scattering model [GBGP02], which can be regarded as an extension of the separable Rayleigh model: The receive correlation matrix is modeled as a random matrix. In this case $H_k = R_{Rk}^{1/2} \tilde{H}_{Rk} R_{Sk}^{1/2} \tilde{H}_{Sk} R_{Tk}^{1/2}$, with independent standard normal-distributed matrices $\tilde{H}_{Rk} \in \mathbb{C}^{n_R \times n_{Sk}}$ and $\tilde{H}_{Sk} \in \mathbb{C}^{n_{Sk} \times n_T}$, and an intermediate scattering correlation matrix $R_{Sk} \in \mathbb{C}^{n_{Sk} \times n_{Sk}}$. For the use of single-user MIMO channel models in a multi-user setting, the general channel model assumes statistically independent channels $H_1, \ldots, H_K$, which is natural if the users are geographically separated. Similar generalizations of separable Rayleigh or double-scattering models to the multi-user uplink are also considered in [SU07, BLD09, JG04, LJGM10, HCD11]. We will initially focus on the general channel model, and exemplify key points using the transmit-correlated Rayleigh model. Section 6.5 is, however, restricted to the transmit-correlated Rayleigh model, since it exploits the analytical tractability of the model.

Regarding channel information, we assume that the BS has perfect CSI: at any given time, the realization of $H = [H_1 \ldots H_K]$ is known. CSI is a precondition for using ZF equalization, which also requires the knowledge of the remaining system parameters: the precoders $P_1, \ldots, P_K$ and the order $\Pi$ in which the subchannels in the system are decoded. Assuming a transmission scenario that rules out precoding based on perfect CSI, we shall henceforth focus on the design of these parameters based on long-term, statistical CSI, which is also assumed to be perfectly known at the BS. With optimization carried out at the BS, we assume that optimized precoders can be distributed to the users over a downlink channel.

### 6.1.3 System-Related Notation

For the analysis of ZF-DF equalization with an arbitrary decoding order $\Pi$, it will be convenient to slightly generalize the QL decomposition of an
6.1. SYSTEM MODEL

Table 6.1: Notation for the decoding order.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Pi$</td>
<td>$n_L \times n_L$</td>
<td>Decoding order (permutation matrix)</td>
</tr>
<tr>
<td>$\pi$</td>
<td>$n_L \times 1$</td>
<td>Decoding order (permutation vector)</td>
</tr>
<tr>
<td>$\Pi^{(k)}$</td>
<td>$n_{Lk} \times n_{Lk}$</td>
<td>Intrinsic decoding order of user $k$</td>
</tr>
<tr>
<td>$\pi_k$</td>
<td>$n_{Lk} \times 1$</td>
<td>Decoding order of user $k$</td>
</tr>
</tbody>
</table>

arbitrary complex tall matrix $T$, as well as the corresponding Cholesky elements used in preceding chapters:

**Definition 6.1.** Given $T \in \mathbb{C}^{M \times N}$ with $M \geq N$, and an $N \times N$ permutation matrix $\Pi$, the matrices $Q \in \mathbb{C}^{M \times N}$ and $L \in \mathbb{C}^{N \times N}$ comprise a $\Pi$-ordered QL decomposition of $T$ if $QL = T$, $Q$ has orthonormal columns, and $\Pi L \Pi^T$ is lower triangular. The corresponding vector of Cholesky elements $l_{\Pi}^2(T^H T) \in \mathbb{R}^N$ is defined by the diagonal of $L$ as $l_{\Pi}^2(T^H T) = [l_{11}^2 \ldots l_{NN}^2]^T$.

This definition enables a compact expression for the system SNR vector $\gamma$ in Section 6.1.1 as

$$\gamma = l_{\Pi}^2(P^H H^H H P).$$

(6.2)

In the following, we shall use an alternative representation of the decoding order: The permutation vector $\pi \in S_\pi$ has unique integer elements from the set $\{1, \ldots, n_L\}$, and $S_\pi$ denotes the set of $n_L!$ possible permutation vectors of length $n_L$. This vector can be partitioned as $\pi = [\pi_1^T \ldots \pi_K^T]^T$, where the $n_{Lk}$-component vector $\pi_k = [\pi_{k,1} \ldots \pi_{k,n_{Lk}}]^T$ contains the decoding order for user $k$’s substreams in the following sense: Substream $n$ of user $k$ is the $\pi_k$th substream in the system to be decoded; $\pi$ is related to $\Pi$ as $\pi = \Pi^T \pi_\uparrow$ for the identity permutation $\pi_\uparrow = [1 \ldots n_L]^T$. Similarly, there is an $n_{Lk} \times n_{Lk}$ permutation matrix $\Pi^{(k)}$ associated with $\pi_k$ in the sense that $\pi_k = \Pi^{(k)} T \pi_{k \uparrow}$, where $\pi_{k \uparrow}$ is the increasing rearrangement of $\pi_k$. We refer to $\Pi^{(k)}$ as the *intrinsic decoding order* of user $k$, since $\Pi^{(k)} s_k$ is decoded in order from the first component to the last (but other users’ substreams may be decoded in between). The notation related to the decoding order is summarized in Table 6.1. For the example given in Figure 6.2, $\pi = [2 4 3 1]^T$, and

$$\pi_1 = \begin{bmatrix} 2 \\ 4 \end{bmatrix}, \quad \pi_2 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \quad \Pi^{(1)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \Pi^{(2)} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. $$
6.2 User Utility Functions

This section introduces the user perspective on performance of communication in the uplink. We shall make use of the notion of a user utility: a real-valued quantity $u_k$ reflecting the performance of user $k$. Using the convention that an increase in utility leads to better performance, a typical example of a user utility is the maximum achievable data rate of reliable communication. Other common performance measures, such as a detection-error probability, mean square error (MSE), probability of outage, etc., qualify as negative utilities. From the user perspective, system performance in the uplink is characterized by the utility vector $u = [u_1 \ldots u_K]^T$.

The utility of user $k$ naturally depends on the quality of the $n_{L_k}$ subchannels utilized for spatial multiplexing; $u_k$ depends on the joint distribution of user $k$’s subchannel SNRs $\gamma_k = [\gamma_{k,1} \ldots \gamma_{k,n_{L_k}}]^T$. When defined in this way, each user utility is a function of the parameters $P_1, \ldots, P_K$ and $\pi$, which we shall refer to as a communication mode; this notion captures all system-design parameters to be optimized in the following.\(^1\) A transmit-power budget $P_{\text{max}}$ for each user is also made explicit in the following definition of a pure communication mode (as opposed to a mixed communication mode, which will be defined in Section 6.3).

**Definition 6.2.** A pure communication mode $M$ is a tuple on the form $M = (P_1, \ldots, P_K, \pi)$. $M$ is called feasible if $\text{tr} (P_k^H P_k) \leq P_{\text{max}}$ for each $k = 1, \ldots, K$.

We shall proceed to investigate in detail how the design parameters affect general user utilities. The following theorem characterizes the SNRs for all subchannels in the MIMO uplink.

**Theorem 6.1.** Let $\Pi$ be a decoding order with corresponding intrinsic decoding orders $\Pi^{(1)}, \ldots, \Pi^{(K)}$. For each user $k$, let $x_k = l_{\Pi^{(k)}}^2 (P_k^H R_{tk}^T P_k)$, and form $x = [x_1^T \ldots x_K^T]^T$. Under the general channel model assuming $H_k = \tilde{H}_k R_{tk}^{1/2}$, let $\tilde{H}_{-k}$ denote the submatrix of $\tilde{H}_k$ consisting of its first $n_{L_k}$ columns. Forming $\tilde{H}_{-} = [\tilde{H}_{-1} \ldots \tilde{H}_{-K}]$, the system SNR vector in (6.2) is distributed as $\gamma \sim l_{\Pi}^2 (\tilde{H}^H \tilde{H}_{-}) \odot x$. The joint distribution of user $k$’s SNRs $\gamma_k = [\gamma_{k,1} \ldots \gamma_{k,n_{L_k}}]^T$ is parameterized only by $P_k$ and $\pi$.

**Proof.** See Appendix 6.A.1. \(\square\)

A crucial observation in Theorem 6.1 is the fact that the choice of precoder for a certain user does not affect the other users’ utilities. This result

---

\(^1\)Note that the number of subchannels $n_{L_k}$ available to user $k$ may also affect its performance. Here we assume that $n_{L1}, \ldots, n_{LK}$ are fixed, and focus on optimizing $P_1, \ldots, P_K$ and $\pi$. If this problem can be efficiently solved, the outer optimization of $n_{L1}, \ldots, n_{LK}$ subject to $\sum_{k=1}^K n_{L_k} \leq n_R$ can also be addressed, but lies beyond the scope here.
holds under the general channel model in Section 6.1.2, and obviously has implications for precoder optimization that we will exploit in the following. Glancing again at Theorem 6.1, it is natural to express the utility of user \( k \) as a function \( u_k = u_k(x_k, \pi) \). \( x_k \) can be interpreted as the power allocated to the subchannels. Since increasing \( x_{n_k} \) leads to a more favorable distribution of user \( k \)'s SNR vector \( \gamma_k \), we assume that \( u_k(x_k, \pi) \) is increasing in each component of \( x_k \). Motivated by our findings up to this point, we define a normal utility function as follows.

**Definition 6.3.** A normal utility function \( u(M) \) can be written on the form \( u(M) = [u_1(x_1, \pi) \ldots u_K(x_K, \pi)]^T \) where \( x_k = I_{P_H}^2(P_k^H R_{Tk} P_k) \). Each \( u_k(x_k, \pi) \) is continuous and component-wise increasing in \( x_k \) for \( x_k \geq 0 \) and any \( \pi \in S_\pi \).

As a corollary of Theorem 6.1, we present the joint SNR distribution in the case of transmit-correlated Rayleigh fading. Here, \( \gamma \sim \Gamma(\alpha, \theta) \) denotes the gamma distribution with shape parameter \( \alpha \) and scale parameter \( \theta \), corresponding to the probability density function \( p_\gamma(z) = z^{\alpha-1}e^{-z/\theta}/[(\alpha - 1)!\theta^\alpha] \).

**Corollary 6.1.** Under the transmit-correlated Rayleigh model, the subchannel SNRs \( \{\gamma_{k,n}\}_{k,n} \) are statistically independent and distributed as \( \gamma_{k,n} \sim \Gamma(n_R - n_L + \pi_{k,n}, x_{k,n}) \), where \( x_k = [x_{k,1} \ldots x_{k,n_{1,k}}]^T = I_{P_H}^2(P_k^H R_{Tk} P_k) \) for \( k = 1, \ldots, K \).

**Proof.** With \( \text{vec}(\hat{H}_-) \sim \mathcal{CN}(0, I) \), the components of \( I_{P_H}^2(\hat{H}_-^H \hat{H}_-) \) are statistically independent and distributed as \( l_{1,m}(\hat{H}_-^H \hat{H}_-) \sim \Gamma(n_R - n_L + m, 1) \) (see, e.g., [TV04, Lemma 2.1]). By definition, \( I_{P_H}^2(\hat{H}_-^H \hat{H}_-) = \Pi^T l_{1}^2(\Pi \hat{H}_-^H \hat{H}_- \Pi^T) \), and by the unitary invariance of \( \hat{H}_- \), \( I_{P_H}^2(\Pi \hat{H}_-^H \hat{H}_- \Pi^T) \sim l_{1}^2(\hat{H}_-^H \hat{H}_-) \). Theorem 6.1 then states that \( \gamma \sim \Pi^T l_{1}^2(\hat{H}_-^H \hat{H}_-) \odot x \), and the corollary follows.

Corollary 6.1 has an important implication regarding the amount of signalling required to implement joint precoder–decoding-order optimization in transmit-correlated Rayleigh fading. Once the optimal communication mode \( M^* = (P_1^*, \ldots, P_K^*, \pi^*) \), has been established at the BS, the optimal precoder \( P_k^* \) can be constructed at user \( k \), using only local channel information in terms of \( R_{Tk} \) along with the decoding order \( \pi_k^* \) for its subchannels. The latter can be obtained with only a few bits \( (\log_2(n_L!/(n_L - n_{1,k})!)) \) of feedback from the BS to the user. \( R_{Tk} \) can often be estimated locally at the user without any additional feedback; in a time-division duplexing system, \( R_{Tk} \) can be obtained using channel reciprocity, while in a frequency-division duplexing system, an estimate of \( R_{Tk} \) can be similarly obtained after a transformation compensating for the frequency offset [BO01].
6.2.1 Examples of User Utility Functions

User utility functions can be composed in a similar manner as in the single-user case, which was dealt with in detail in Section 3.2.2. Here, we shall only review a few examples that will be important for the following discussion.

- **Achievable data rate:** The achievable sum-rate for user $k$ in fast fading becomes

  \[ u_k^{\text{Rate}}(x_k, \pi) = \sum_{n=1}^{n_{L_k}} \mathbb{E} \left[ \log_2 (1 + \gamma_{k,n}) \right]. \quad (6.3) \]

- **Probability of successful detection:** Using a square $M_{k,n}$-QAM constellation on subchannel $n$ of user $k$, a useful utility is the probability of successfully detecting all symbols, which is the complementary joint symbol error probability (JEP). Under the simplifying assumption of no error propagation from other users in the ZF-DF equalization process, this utility can be expressed as

  \[ u_k^{1\text{-JEP}}(x_k, \pi) = \mathbb{E} \left[ \prod_{n=1}^{n_{L_k}} \left( 1 - f_{M_{k,n}}^{\text{SEP}}(\gamma_{k,n}) \right) \right], \quad (6.4) \]

  where $f_{M_{k,n}}^{\text{SEP}}(\gamma)$ denotes the probability of error with SNR $\gamma$ as defined in (3.13).

- **Weighted MSE:** An MSE-based utility using positive weights $\alpha_{k,1}, \ldots, \alpha_{k,n_{L_k}}$ is given by the weighted sum

  \[ u_k^{-\text{wMSE}}(x_k, \pi) = -\sum_{n=1}^{n_{L_k}} \alpha_{k,n} \mathbb{E} \left[ \gamma_{k,n}^{-1} \right]. \quad (6.5) \]

  As argued in [BPO09], it is appropriate to use rate-normalizing weights $\alpha_{k,n}^{\text{RN}} = 2R_{k,n} - 1$, where $R_{k,n}$ denotes the intended data rate on the subchannel.

6.3 Utility Regions

This section elaborates on the user perspective on performance in the uplink by focusing on the utility region: the set of real-valued $K$-tuples of user utilities that can be achieved using feasible pure communication modes (see Def. 6.2). We shall distinguish between two kinds of utility regions: one based on transmission using a single mode, and one involving multiple modes. The latter can be realized using time-sharing. In a time-sharing cycle, each of the involved communication modes is allocated time according to a time-sharing coefficient. Alternatively, multi-mode transmission
can be implemented by choosing a communication mode at random based on some probability distribution. It is possible to extend the notion of a communication mode to incorporate multi-mode transmission in a unified way; we define a *mixed communication mode* to be a random variable with a finite number of pure communication modes as outcomes. Time-sharing is included by letting the probability associated with a particular pure mode represent its time-sharing coefficient.

**Definition 6.4.** A *mixed communication mode* $\mathcal{M}$ is a random variable with support in $\{M_1, \ldots, M_I\}$: a finite set of pure communication modes. $\mathcal{M}$ is called *feasible* if every $M_i$ is feasible for $i = 1, \ldots, I$.

We shall formalize the notion of a utility region over an arbitrary set $S$ of communication modes. It is then necessary to define user utilities $u_{\mathcal{M}} \in \mathbb{R}^K$ for a mixed communication mode $\mathcal{M}$. For a normal utility function $u(M)$, it is natural to let $u_{\mathcal{M}} = \mathbb{E}[u(M)] = \sum_{i=1}^{I} \Pr(\mathcal{M} = M_i) u(M_i)$. In the context of time-sharing, this definition represents average performance over the time-sharing cycle for each user, and when a single pure mode $\mathcal{M} = M$ is employed we simply have $u_{\mathcal{M}} = u(M)$.

**Definition 6.5.** For a normal utility function $u(M)$, the *utility region over $S$* is the subset of $\mathbb{R}^K$ that comprises all $u_{\mathcal{M}} = \mathbb{E}[u(M)]$ that can be achieved using a communication mode $\mathcal{M} \in S$.

Definition 6.4 is sufficient for identifying the utility region over mixtures of feasible pure communication modes; it can be deduced from [BV04, Sec. 2.1.4] that there is no restriction in only considering mixed modes based on a finite number of pure modes. The notions of pure and mixed communication modes correspond to similar definitions of *pure and mixed strategies* in game theory [MD06]; a player in a game may employ a mixed strategy by randomly selecting a single, pure strategy based on a probability distribution. Considering the expected utility under a mixed communication mode is also in line with the game-theoretic view: A rational player would employ a mixed strategy that maximizes its expected utility. We shall, however, not stress an analogy between users in the MIMO uplink and players in a game due to Theorem 6.1; an individual user’s precoding strategy has no impact on the other users’ utilities, and there is hence no competition among users.

We shall approach the characterization of utility regions by fixing the decoding order $\pi$, and focusing on the precoders $P_1, \ldots, P_K$. For any normal utility function $u(M) = [u_1(x_1, \pi) \ldots u_K(x_K, \pi)]^T$, the choice of

---

2However, it is possible to relax the feasibility condition that $\text{tr}(P_k^H P_k) \leq P_{\text{max}}$ is to be fulfilled for each possible outcome of a mixed mode; it may suffice to only require that the transmit-power constraint is fulfilled on average [BLD09]. While such an extension may very well extend the utility region, it represents a somewhat different transmission scenario, and will not be considered here.
precoder $\mathbf{P}_k$ only affects the utility of user $k$ since $x_k = l_{H(\mathbf{P}_k)}^2 (\mathbf{P}_k H R T_k P_k)$. Thus, the maximal utility $\hat{u}_k (\pi)$ of user $k$ over feasible pure modes with decoding order $\pi$ is determined by the optimization problem

$$\text{maximize } \quad u_k (x_k, \pi)$$

subject to

$$x_k = l_{H(\mathbf{P}_k)}^2 (\mathbf{P}_k H R T_k P_k),$$

$$\text{tr} (\mathbf{P}_k H P_k) \leq P_{\text{max}}.$$  \hspace{1cm} (6.6)

Interestingly, (6.6) is in essence a single-user precoder design problem. Such problems have been studied in detail in Chapter 3 for general performance measures, resulting in tractable design formulations. A key point that follows merely from $u_k (x_k, \pi)$ being increasing in $x_k$ is that it is optimal to let the transmit directions (left singular vectors) of $\mathbf{P}_k$ match the eigenvectors of $R T_k$ corresponding to its $n_L k$ largest eigenvalues. Due to our characterization of the SNR distribution in Theorem 6.1, the optimal single-user transmit directions remain the same in the presence of other users. This result holds for any normal utility function under the general channel model for ZF-DF equalization. This complements the view on the same conclusion obtained by [SU09] in the related setting of sum-rate maximization in transmit-correlated Rayleigh fading for minimum-MSE DF equalization.

In transmit-correlated Rayleigh fading, several common choices of $u_k (x_k, \pi)$ have an analytical expression and the problem (6.6) can be solved using convex optimization. For other kinds of fading, $u_k (x_k, \pi)$ can be harder to evaluate. The expressions in Chapter 4 only apply if there is a common receive-correlation matrix for all users. In the general case, it is possible to use Monte-Carlo based methods; for example, [JV08b] provides means to find approximate expressions for subchannel error- and outage probabilities as functions of the subchannel power allocation, using a four-parameter hyperbola model. While the method is straightforward to apply, it comes at the price of being computationally demanding. We shall not further investigate the single-user precoding problem here; in the remainder we make the simplifying assumption that optimization problems on the form (6.6) can be solved. Next we present the utility region $U_{\text{PM}}$ over the set of feasible pure modes $S_{\text{PM}}$.

**Theorem 6.2.** For a normal utility function $u (M)$, let $\bar{u}_k (\pi) = u_k (0, \pi)$ and let $\hat{u}_k (\pi)$ denote the maximal utility of (6.6). The utility region over all feasible pure communication modes $S_{\text{PM}}$ is given by $U_{\text{PM}} = \bigcup_{\pi \in S_{\text{PM}}} U_{\pi_{\text{PM}}}$, where $U_{\pi_{\text{PM}}}$ denotes the hyper-rectangle

$$U_{\pi_{\text{PM}}} = \{ \mathbf{u} \in \mathbb{R}^K : \bar{u}_k (\pi) \leq u_k \leq \hat{u}_k (\pi), \, k = 1, \ldots, K \}.$$  \hspace{1cm} (6.7)

**Proof.** See Appendix 6.A.2. \hfill $\square$
The utility region $U_{\pi}^{PM}$ over feasible pure modes with decoding order $\pi$ is nothing but a hyper-rectangle for normal utility functions. The lower corner point $\hat{u}(\pi) = [\hat{u}_1(\pi) \ldots \hat{u}_K(\pi)]^T$ corresponds to the trivial case of having zero SNR on each subchannel (see Thm. 6.1). For the three examples of user utilities in Section 6.2.1, we have: $\hat{u}_k^{\text{Rate}}(\pi) = 0$ reflecting the fact that the subchannels cannot convey information; $\hat{u}_k^{1-\text{JEP}}(\pi) = 1/(M_k \cdot \ldots \cdot M_{k,n_{lk}})$, which is equal to mere guessing; and $\hat{u}_k^{\text{wMSE}}(\pi) = -\infty$.

The utility region $U^{PM}$ over pure modes is formed by taking the union of fixed-order utility regions $U_{\pi}^{PM}$ over all decoding orders $\pi \in S_{\pi}$. This results in $U^{PM}$ being a polyblock: a finite union of hyper-rectangles. Allowing for mixed communication modes, with an averaging effect on utility as $u_M = \mathbb{E}[u(M)]$, results in convexifying the pure-mode utility region as the following theorem clarifies. Since $U^{PM}$ is a polyblock, the mixed-mode utility region $U^{MM}$ becomes a convex polytope.

**Theorem 6.3.** For a normal utility function $u(M)$, the utility region over all feasible mixed communication modes $S^{MM}$ is given by $U^{MM} = \text{conv}(U^{PM})$.

**Proof.** See Appendix 6.3. \qed

### 6.3.1 Pareto Efficiency

From a general multi-objective optimization perspective, there is only a single point of practical interest in the set $U_{\pi}^{PM}$: the component-wise precoder-optimized utility vector $\hat{u}(\pi) = [\hat{u}_1(\pi) \ldots \hat{u}_K(\pi)]^T$. The reason is that $\hat{u}(\pi) \geq u$ for any $u \in U^{PM}$. This reflects the fact that for a given decoding order, each user’s utility is maximized by optimizing its precoder $P_k$. Things become more interesting for the utility regions $U^{PM}$ and $U^{MM}$, since there is no longer a single operating point $\hat{u}$ satisfying $\hat{u} \geq u$ for any $u$ in the region. However, points in the utility region that are Pareto efficient are of immediate interest. At such points, the utility of any user can only be increased at the expense of other users’ utilities. More formally, a utility vector $\hat{u} \in U$ is called Pareto efficient in $U$ if there is no other $u \in U$ such that $\hat{u} \leq u$ component-wise. We shall make use of the set of optimized utility vectors denoted by

$$V = \{\hat{u}(\pi) : \pi \in S_{\pi}\}.$$ \hspace{1cm} (6.8)

Any Pareto-efficient point in $U^{PM}$ belongs to the set $V$, and any Pareto-efficient point in $U^{MM}$ belongs to the convex polytope $\text{conv}(V)$. This is a

---

*3Due to the singularity of $u_k^{wMSE}(x_k, \pi)$ at $x_k = 0$, MSE-based utilities are not strictly “normal” according to Def. 6.3. However, the results for normal utility functions apply with trivial modifications.*
CHAPTER 6. PRECODING AND ORDERING IN THE UPLINK

Figure 6.3: Illustration of utility regions in a symmetric two-user uplink with \( n_{T1} = n_{T2} = 2, n_R = 6, P_{\text{max}1} = P_{\text{max}2} = 10 \text{ dB}\) and spatially uncorrelated Rayleigh-fading MIMO channels. Each user transmits 6 bits using \( n_{L1} = n_{L2} = 2 \) substreams (one 16-QAM and one 4-QAM).

consequence of the following basic lemma, which will be used as an auxiliary result in the next section.

Lemma 6.1. For any \( u \in U^{PM} \) (\( u \in U^{MM} \)) there is a \( \hat{u} \in \mathcal{V} \) (\( \hat{u} \in \text{conv}(\mathcal{V}) \)) such that \( u \leq \hat{u} \).

Proof. See Appendix 6.A.4. \( \square \)

The different utility regions \( U^{PM}_\pi, U^{PM}, \) and \( U^{MM} \), as well as the set \( \mathcal{V} \), are illustrated in Figure 6.3 for the simple case of a symmetric two-user uplink. Here, user utility is measured in terms of the negative total bit-error probability (BEP) for each user (assuming no error propagation). With two subchannels per user, there are \( |\mathcal{V}| = 24 \) possible decoding orders, and 6 pareto-efficient utilities in \( U^{PM} \) (corresponding to the 6 possible partitions of the set \( \{1, 2, 3, 4\} \) of placements in the decoding order into two two-element sets).

6.4 Maximizing System Utility

Having investigated performance from a user perspective, we shift the focus to the related matter of optimizing system performance. Instead of considering user utilities as points in \( \mathbb{R}^K \), with one component for each
6.4. MAXIMIZING SYSTEM UTILITY

user, we shall consider a scalar system performance measure $U \in \mathbb{R}$. Here, we assume that system utility is based on user utilities: We let $U_M = \mathcal{H}(u(M))$, where $\mathcal{H}$ denotes an aggregate utility function, which is assumed to be increasing in each argument. As an example, consider combining the subchannel sum rate $u_k^{\text{Rate}}(x_k, \pi)$ in (6.3) with the aggregate function $\mathcal{H}(u) = \min_k u_k$ [SB04]. This results in a system utility function targeting user fairness as

$$U_{\text{min-rate}} = \min_k u_k^{\text{Rate}}(x_k, \pi).$$

(6.9)

As a second example, consider the case of uncoded transmission with a user utility $u_k^{1-\text{JEP}}(x_k, \pi)$ as in (6.4), assuming transmit-correlated Rayleigh fading. Defining system utility as the probability of successfully detecting all symbols in the system amounts to

$$U_{1-\text{JEP}}^M = \prod_{k=1}^{K} u_k^{1-\text{JEP}}(x_k, \pi),$$

(6.10)

where $\mathcal{H}(u) = \prod_k u_k$.

Provided $u(M)$ and $\mathcal{H}(u)$, it is still not obvious how to represent system performance for a mixed communication mode. One possibility is to apply $\mathcal{H}(u)$ on the mixed-mode performance region as $U_M = \mathcal{H}(u_M)$ with $u_M = \mathbb{E}[u(M)] \in U_{\text{MM}}$. This is a suitable definition for the example provided in (6.9): Each user’s data rate is averaged over the time-sharing cycle, prior to being compared via the aggregate function $\mathcal{H}(u) = \min_k u_k$. Another possibility is to apply $\mathcal{H}(u)$ and the expectation the other way around as $U_M = \mathbb{E}[\mathcal{H}(u(M))]$. This definition is more appropriate when generalizing the second example in (6.10) to mixed modes. If multiple pure modes are employed, then the average probability of successful decoding over the time-sharing cycle is on the form $\mathbb{E}[\mathcal{H}(u(M))]$.

We shall approach the task of optimizing system performance by investigating the optimality of using pure communication modes. Are there system utilities $U_M$ for which the optimal pure mode performs as well as any mixed mode? This question is answered by the following theorem.

**Theorem 6.4**. Assume that $\mathcal{H}(u)$ is component-wise increasing, $u(M)$ is a normal utility function, and that either

- $U_M = \mathbb{E}[\mathcal{H}(u(M))]$;
- or $U_M = \mathcal{H}(\mathbb{E}[u(M)])$ and $\mathcal{H}(u)$ is convex.

Then there exists a feasible pure mode $M^*$ such that $U_{M^*} \geq U_M$ for any feasible mixed mode $M$.

**Proof.** See Appendix 6.A.5. $\square$
Motivated by Theorem 6.4, we shall take a closer look at the optimization of a single pure communication mode. We shall also consider mixed-mode optimization for the relevant case when \( U_M = \mathcal{H}(\mathbb{E}[u(M)]) \) and \( \mathcal{H}(u) \) is not convex. Both problems can be regarded as the optimization of an aggregate function \( \mathcal{H}(u) \) over a utility region \( U \), or \( \max_{u \in U} \mathcal{H}(u) \), where \( U = U^{PM} \) in the single-mode case, and \( U = U^{MM} \) in the multi-mode case.

Under the assumption that \( \mathcal{H}(u) \) is increasing in \( u \), it is evident that Pareto-efficient utilities are of special interest; if \( u^* \in U \) is not Pareto efficient then there exists a \( \tilde{u} \in U \) such that \( \mathcal{H}(u^*) \leq \mathcal{H}(\tilde{u}) \). For the maximization of \( \mathcal{H}(u) \) over the pure-mode utility region \( U^{PM} \), the conclusion of Lemma 6.1 that any Pareto-efficient utility belongs to the finite set \( V \) in (6.8) suggests that a maximizer can be obtained by an exhaustive search.

The general mixed-mode design problem can be similarly addressed. Aided again by Lemma 6.1, we conclude that it is not necessary to optimize over the entire set \( U \in U^{MM} \); there is an optimal utility vector in the set \( \text{conv}(V) \). There is a straightforward method to optimize over the bounded, convex polytope \( \text{conv}(V) \). Forming the \( K \times n L! \) matrix \( V \), with the points in \( V \) as columns, any \( u \in \text{conv}(V) \) can be written as \( u = V \theta \) using an \( n L! \)-vector \( \theta \) with non-negative components adding up to one. This results in the following optimization problem.

\[
\begin{align*}
\text{maximize} & \quad \mathcal{H}(u) \\
\text{subject to} & \quad u = V \theta, \\
& \quad \sum_{i=1}^{n L!} \theta_i = 1, \quad \theta_j \geq 0, \quad j = 1, \ldots, n L!.
\end{align*}
\]

This problem is a convex optimization problem provided that \( \mathcal{H}(u) \) is concave. In fact, the problem can be written as a linear program when optimizing user fairness through \( \mathcal{H}(u) = \min_k u_k \). Once the optimal \( u^* \) and \( \theta^* \) have been found, the corresponding optimal mixed mode uses \( \theta^* \) as multi-mode probabilities associated with the precoder-optimized utilities in \( V \).

The computational complexity associated with both pure- and mixed-mode optimization as described in this section is prohibitive. Both cases rely on the complete characterization of \( V \), which in general requires solving

---

4As an alternative to (6.11), it is possible to utilize the facet description of \( \text{conv}(V) \), which is on the form \( Au \leq a \). Finding the facet description is commonly termed the facet enumeration problem and is well known in computational geometry [GO04, Ch. 22]. A triangulation of the boundary of \( \text{conv}(V) \) is also necessary in order to associate the optimal \( u \) with a mixed mode.
6.5 Effcient Single-Mode Selection

Focusing on transmission using a single pure communication mode is well motivated, not least from a practical viewpoint. It is easier to implement and requires less feedback from the BS to the users. We have also concluded in Theorem 6.4 that single-mode transmission is optimal in many cases. We shall state the design problem of optimizing system utility $U_M = \mathcal{H}(u(M))$ over the set of feasible pure modes. Abusing notation slightly, we write

$$U(x, \pi) = \mathcal{H}(u_1(x_1, \pi), \ldots, u_K(x_K, \pi))$$

(6.12)

to stress that system utility depends on the power-allocation vectors collected in $x = [x_1^T \ldots x_K^T]^T$, with $x_k = l_{\Pi(k)}^2(P_k^H R_{Tk} P_k)$, as well as on the decoding order $\pi$. The single-mode optimization problem $\max_{u \in U^{PM}} \mathcal{H}(u)$ can then be written as

\[
\begin{align*}
\maximize_{P_1, \ldots, P_K, x, \pi} & \quad U(x, \pi) \\
\text{subject to} & \quad x_k = l_{\Pi(k)}^2(P_k^H R_{Tk} P_k), \quad k = 1, \ldots, K, \\
& \quad \text{tr}(P_k^H P_k) \leq P_{\max}, \quad k = 1, \ldots, K, \\
& \quad \pi \in S_{\pi}.
\end{align*}
\]

(6.13)

We shall explore the technique of alternating optimization to find a local maximizer of (6.13). It is natural to partition the optimization variables into $x$ and $\pi$. This allows focusing on the two sub-problems of optimal pre-coding (keeping $\pi$ fixed) and optimal decoding ordering (keeping $x$ fixed). If both of these sub-problems can be solved, it is possible to perform alternating optimization to iterate towards a local maximum. Instead of performing an exhaustive search over the entire set of precoder-optimized utilities $\mathcal{V}$ in (6.8), an alternating-optimization based method traverses a finite number of points in $\mathcal{V}$ until convergence. This method has the potential to drastically reduce the number of points in $\mathcal{V}$ that need to be computed.

\footnote{For channel models with certain symmetry, such as the transmit-correlated Rayleigh model, only $\sum_{k=1}^K \frac{n_k l!}{(n_l - n_{L_k})!}$ precoding problems need to be solved, since the utility of user $k$ depends only on $\pi_k$ instead of the entire $\pi$.}
The precoding sub-problem, assuming a fixed decoding order \( \pi \) in (6.13), leaves the problem of jointly optimizing all users’ precoding matrices based on all users’ statistical CSI. Assuming a system utility function (6.12) with \( \mathcal{H}(u) \) being component-wise increasing in \( u \), (6.13) becomes trivially parallelizable in the precoders. The precoder of user \( k \) is given by solving a single-user precoding problem as described in Chapter 3, taking the form

\[
\begin{align*}
\text{maximize} & \quad u_k (x_k, \pi) \\
\text{subject to} & \quad \log x_k \leq \log p_k + \log \lambda_{T_k}, \\
& \quad \sum_{n=1}^{n_{L_k}} p_{k,n} \leq P_{\max},
\end{align*}
\]

where \( p_k \in \mathbb{R}^{n_{L_k}} \) denotes the eigenvalues of \( P^H_k P_k \), and \( \lambda_{T_k} \) is the decreasing vector containing the \( n_{L_k} \) largest eigenvalues of \( R_{T_k} \). It is then possible to directly optimize \( x = [x_1^T \ldots x_K^T]^T \), given a fixed decoding order \( \pi \). The other sub-problem of finding an optimal decoding order is investigated next. Details on the alternating optimization approach are provided in Section 6.5.2.

### 6.5.1 Optimal Decoding Ordering

This section assumes that fixed subchannel power allocations \( x_1, \ldots, x_K \) have been selected. The ordering problem then amounts to associating each data substream \( s_{k,n} \) in the system with a place \( \pi_{k,n} \in \{1, \ldots, n_{L} \} \) in the decoding order (see Fig. 6.2). The mapping between the sets \( \{s_{k,n}\}_{k,n} \) and \( \{1, \ldots, n_{L}\} \) is bijective: Each place in the decoding order should be assigned to one, and only one, substream, and vice versa. The assignment we are seeking is the one that maximizes the utility of the system, i.e., the maximizer of

\[
\begin{align*}
\text{maximize} & \quad U(x, \pi) \\
\text{subject to} & \quad \pi \in \mathcal{S}_n
\end{align*}
\]

This optimization problem falls into the category of assignment problems (AP) [Pen07, BDM09], which is a branch of combinatorial optimization. A solution can be found by performing an exhaustive search over the \( n_L! \) possible candidates of \( \pi \), but with a computational complexity growing prohibitively fast with \( n_L \). Therefore, we shall dedicate the remainder of this section to explore means to circumvent such a computationally costly approach. In particular, we apply results from the theory of linear assignment problems, which have been rigorously investigated in the literature [BDM09]. First, we present a general classification of decoding-ordering problems into different classes of linear assignment problems. This
admits the ordering problem to be solved using efficient, standard
algorithms in many cases. Second, we identify relevant cases when the ordering
problem can be solved in closed form.

We shall focus here on the case when each subchannel has a corre-
sponding utility function in the following sense: $u_{k,n}(x_{k,n}, \pi_{k,n})$ denotes
the utility of subchannel $n$ of user $k$ as a function of the power $x_{k,n}$ al-
located to the subchannel as well as its place $\pi_{k,n}$ in the decoding order.
It was concluded in Section 6.2.1 that this is indeed the case for several
common performance measures under transmit-correlated Rayleigh fading.
This generalizes to any multi-user channel model rendering $\tilde{H}$ in Theorem
6.1 statistically invariant to column permutations (e.g., the somewhat arti-
ficial case of separable Rayleigh fading with a common receive correlation
matrix for the users [BLD09]). Under the general channel model, however,
the utility of a certain subchannel normally depends not only on its own
place in the decoding order, but rather on the entire decoding order; this
can be deduced from Theorem 6.1.

Characterizing an AP as linear relies on the construction of a utility
matrix $Z$: Given an $n_L \times n_L$ utility matrix $Z$, a linear assignment problem
aims to match the rows of $Z$ to different columns according to a specified
objective [BDM09]. A row-column assignment is represented by a permu-
tation vector $\pi$ in the sense that row $n$ is matched to column $\pi_n$. For the
decoding ordering problem, we shall compose an $n_L \times n_L$ utility matrix $Z$
by vertically concatenating user-utility matrices $Z_1, \ldots, Z_K$ of dimension
$Z_k \in \mathbb{R}^{n_L \times n_L}$. The possible utility values for substream $n$ of user $k$
are gathered in the $n^{th}$ row of $Z_k$. Hence, let the element on row $n$ and column
$l$ of $Z_k$ be given by $[Z_k]_{nl} = u_{k,n}(x_{k,n}, l)$. Then the utility matrix we shall
make use of is given by $Z = [Z_1^T \ldots Z_K^T]^T$.

**The Linear Sum Assignment Problem**

The most classical AP, the linear sum assignment problem (LSAP), arises
when maximizing the sum utility

$$\text{maximize } \sum_{m=1}^{n_L} z_{m, \pi_m}.$$  \hspace{1cm} (6.16)

We end up in this problem formulation when all subchannel utilities in
the system are added up together: user utilities are sums of subchannel
utilities, $u_k(x_k, \pi) = \sum_n u_{k,n}(x_{k,n}, \pi_{k,n})$, and system utility is a sum of
user utilities $U(x, \pi) = \sum_k u_k(x_k, \pi)$. This formulation directly applies to
many relevant system utilities, for example a weighted sum of subchannel
MSEs, bit or symbol error probabilities, the achievable data rate, etc. It
also applies to the product of subchannel utilities, such as $U^{1-\text{JEP}}(x, \pi)$
in (6.10), by optimizing $\log(U(x, \pi))$ instead, which is equivalent. The
LSAP can be reformulated as a linear program, which enables a primal-dual characterization of the problem. Several algorithms of different kind exploit the primal and/or dual formulation, solving the LSAP in $O(n_1^3)$ time. We refer to [BDM09, Ch. 4] for details on such algorithms, as well as for guidance regarding available software.

### The Linear Bottleneck Assignment Problem

Another classical AP maximizes the minimum utility in the system

$$\max_{\pi \in \mathcal{S}_n} \min_{m=1,\ldots,n_L} z_{m,\pi_m}. \quad (6.17)$$

In this formulation, system performance is completely determined by the performance of the weakest subchannel in the system. In terms of user and system utilities, the linear bottleneck assignment problem (LBAP) formulation applies when $u_k(x_k, \pi) = \min_n u_{k,n}(x_{k,n}, \pi_{k,n})$ and $U(x, \pi) = \min_k u_k(x_k, \pi)$. The LBAP can be solved in $O(n_L^2 \sqrt{n_L / \log n_L})$ time [BDM09].

### Categorized Linear Assignment Problems

There are two useful formulations that are hybrids of the previous ones. Firstly, when user fairness is in focus, it is reasonable to maximize $U(x, \pi) = \min_k u_k(x_k, \pi)$. Letting each user utility be a sum of subchannel utilities, $u_k(x_k, \pi) = \sum_n u_{k,n}(x_{k,n}, \pi_{k,n})$, leads to the AP

$$\max_{\pi \in \mathcal{S}_n} \min_{k=1,\ldots,K} \sum_{m=N_k+1}^{N_{k+1}} z_{m,\pi_m}, \quad (6.18)$$

where $N_k = \sum_{j=1}^{k-1} n_L k$, and rows of $Z$ in the range $N_k + 1 \leq m \leq N_{k+1}$ represent the utility of user $k$. Reversely, assuming $U(x, \pi) = \sum_k u_k(x_k, \pi)$ and $u_k(x_k, \pi) = \min_n u_{k,n}(x_{k,n}, \pi_{k,n})$ renders

$$\max_{\pi \in \mathcal{S}_n} \sum_{k=1}^{K} \min_{N_k+1 \leq m \leq N_{k+1}} z_{m,\pi_m}. \quad (6.19)$$

The two categorized APs are unfortunately in general NP-hard [BDM09, Ch. 5.5.2]. Efficient heuristic algorithms have been suggested in [PA93]. Note that the two categorized APs (6.18) and (6.19) reduce to an LBAP and an LSAP, respectively, when each user only utilizes a single subchannel.

The characterization of decoding-ordering problems as linear APs is particularly fruitful since it gives access to an algorithmic framework to solve the problem. However, we shall exemplify a few interesting instances.
of the decoding-ordering problem that can be solved directly, without re-sorting to AP algorithms. As it turns out, it is possible to reuse the general result in Theorem 3.2, which was originally derived to simplify single-user precoding problems. Next, we show how this result can be applied in the new context of optimal decoding ordering.

Closed-Form Solutions

Up to this point, we have allowed each subchannel in the system to have a unique rule to associate a power allocation $x$ and a place $\pi$ in the decoding order with a real-valued utility, as indicated by the subscript in the notation $u_{k,n}(x, \pi)$. This may be appropriate when different coding schemes are used on the subchannels, or to reflect a priority between subchannels or users. Next, we assume that there is no such distinction between subchannels in the system; all subchannels are evaluated based on the same rule. In the following theorem, two $n_L$-vectors $b$ and $c$ are said to be *similarly ordered* (reversely ordered) if $b_m > b_n$ implies $c_m \geq c_n$ ($c_m \leq c_n$).

**Theorem 6.5.** Consider the optimization problem

$$\max_{\pi \in \mathcal{S}_n} U(x, \pi) = \mathcal{F}(u(x_{1,1}, \pi_{1,1}), \ldots, u(x_{K,n_L,K}, \pi_{K,n_L,K})),$$

(6.20)

where $\mathcal{F} : \mathbb{R}^{n_L} \to \mathbb{R}$ is component-wise increasing, and

$$u(x_{k,n}, \pi_{k,n}) = \mathbb{E}[f(\gamma_{k,n})],$$

under the joint distribution of $\{\gamma_{k,n}\}$ in Corollary 6.1 for an increasing function $f : \mathbb{R}_+ \to \mathbb{R}$. If $\mathcal{F}$ is *Schur-concave* and each $u(e^{\bar{x}}, \pi)$ is *concave* in $\bar{x}$, there is a maximizer $\pi^*$ of (6.20) such that

- $\pi^*$ and $x$ are reversely ordered.

If, on the other hand, $\mathcal{F}$ is *Schur-convex* and each $u(e^{\bar{x}}, \pi)$ is *convex* in $\bar{x}$, there is a maximizer $\pi^*$ of (6.20) such that

- $\pi^*$ and $x$ are similarly ordered.

**Proof.** The proof of this theorem is omitted, since it is a mere reformulation of Theorem 3.2.

Regarding Schur-convex/concave functions, we refer to the discussion in Section 3.2. Here we simply note that two examples of Schur-concave functions are $\mathcal{F}(z) = \sum_i z_i$ and $\mathcal{F}(z) = \min_i z_i$, which are precisely the objectives in the LSAP and LBAP, respectively. The function $\sum_i w_i$ is also Schur-convex.

When Theorem 6.5 applies, the optimal decoding order is given by the operation of sorting the $n_L$-vector $x$, which can be performed in
\( O(n_L \log n_L) \) time. It was verified in Section 3.2 that the expected MSE, bit and symbol error probabilities, as well as the outage probability are all convex functions in \( \bar{x} \). Regarding these as negative utilities gives a concave \( u(e^{\bar{x}}, \pi) \) in Theorem 6.5. Combining such utility functions with a Schur-concave \( \mathcal{F}(z) \), as in the case of the LSAP and the LBAP, produces an optimal decoding order \( \pi^* \) that compensates for an unfair distribution of the subchannel power allocations represented by \( x \). In this case, a subchannel with less power \( x_{k,n} \) is placed later in the decoding order. This corresponds to an ordering gain, since subchannels decoded later benefit more from successive interference cancellation. The reverse strategy is optimal when the second part of the theorem applies. It is then optimal to allocate more ordering gain to strong subchannels. This is the case when \( \mathcal{F}(z) \) is Schur-convex, as in the LSAP, in conjunction with, for example, \( u(e^{\bar{x}}, \pi) \) representing the achievable date rate \( \mathbb{E} \left[ \log (1 + \gamma_{k,n}) \right] \) or the expected SNR \( \mathbb{E} \left[ \gamma_{k,n} \right] \).

As an example, we show how Theorem 6.5 applies to MSE-based utility functions. Let \( u^{-\text{MSE}}(x_{k,n}, \pi_{k,n}) \) denote the negative expected MSE on a subchannel, which under transmit-correlated Rayleigh fading becomes (see Sec. 3.2)

\[
\begin{align*}
    u^{-\text{MSE}}(x_{k,n}, \pi_{k,n}) &= \mathbb{E} \left[ -\gamma_{k,n}^{-1} \right] = -x_{k,n}^{-1}(n_R - n_L - 1 + \pi_{k,n})^{-1}.
    \end{align*}
\]  

(6.21)

The function \( f(\gamma) = -\gamma^{-1} \) is increasing, and \( u^{-\text{MSE}}(e^{\bar{x}}, \pi) \) is concave in \( \bar{x} \) for any positive integer \( \pi \) (provided that \( n_R > n_L \)). Theorem 6.5 states that choosing a decoding order \( \pi^* \) such that \( \pi^* \) and \( x \) are reversely ordered is optimal for any increasing and Schur-concave \( \mathcal{F}(z) \). In particular, this is an optimal solution to both the MSE-based LSAP (6.16) as well as the LBAP (6.17).

Theorem 6.5 can be utilized to solve APs for an even larger class of MSE-based utility functions. We can associate a weight \( \alpha_{k,n} > 0 \) with each subchannel MSE as

\[
\begin{align*}
    u^{-\text{wMSE}}_{k,n}(x_{k,n}, \pi_{k,n}; \alpha_{k,n}) &= \alpha_{k,n} u^{-\text{MSE}}(x_{k,n}, \pi_{k,n}),
    \end{align*}
\]  

(6.22)

and obtain the following result.

**Theorem 6.6.** A permutation vector \( \pi^* \) is a maximizer of

\[
\begin{align*}
    \maximize_{\pi \in S_n} \mathcal{U}(x, \pi) = \mathcal{F} \left( \left\{ u^{-\text{wMSE}}_{k,n}(x_{k,n}, \pi_{k,n}; \alpha_{k,n}) \right\}_{k,n} \right),
    \end{align*}
\]

for any Schur-concave and component-wise increasing \( \mathcal{F}(z) \), if

\[\text{6} \text{These functions are convex in } \bar{x} \text{ since they can both be written as a convex mixture of convex functions in } \bar{x}.\]
6.5. EFFICIENT SINGLE-MODE SELECTION

• \( \pi^* \) is reversely ordered to \( \tilde{x} = \left[ \frac{x_{1,1}}{\alpha_{1,1}} \ldots \frac{x_{1,n_{L1}}}{\alpha_{1,n_{L1}}} \frac{x_{2,1}}{\alpha_{2,1}} \ldots \frac{x_{K,n_{Lk}}}{\alpha_{K,n_{Lk}}} \right]^T \).

Proof. We outline the proof as follows. Since \( u_{k,n}^{-\text{wMSE}}(x_{k,n}, \pi_{k,n}; \alpha_{k,n}) = u_{k,n}^{-\text{wMSE}}(x_{k,n}/\alpha_{k,n}, \pi_{k,n}; 1) \), it is possible to reformulate the objective as \( U(\tilde{x}, \pi) \) based on the common subchannel utility function \( u(\tilde{x}_{k,n}, \pi_{k,n}) = u^{-\text{MSE}}(\tilde{x}_{k,n}, \pi_{k,n}) \) for the proposed \( \tilde{x} \). With \( \mathcal{F}(z) \) being increasing and Schur-concave, Theorem 6.5 can be applied on \( U(\tilde{x}, \pi) \), and the result follows.

6.5.2 Alternating Optimization

With efficiently solvable formulations for the two sub-problems, alternating optimization (AO) is a viable approach to select a pure communication mode for transmission. A few comments on convergence, optimality, and initialization of the algorithm are in order. Firstly, AO will naturally converge in a finite number of steps, assuming that the algorithm stops as soon as system utility does not strictly increase. This follows from the fact that the set of possible permutations \( S_\pi \) is finite, and a specific permutation can then only be visited once. However, this does not guarantee that the computational complexity of the algorithm is substantially less than the exhaustive-search approach in Section 6.4. Secondly, the output of the algorithm only satisfies necessary conditions for a global maximizer: The proposed decoding order is optimal for the proposed power allocations, and vice versa. However, we do not necessarily end up with a global maximizer since (6.13) is non-convex and may have multiple such local maxima. The performance of AO is, hence, possibly dependent on how the algorithm is initialized. Improved performance can be achieved by executing a number of instances of AO based on different initializations. In general, this enables selecting the best local maximum out of several candidates. This comes at the expense of a linear increase in computational complexity with respect to the number of AO instances.

We shall leave it to Section 6.6 to evaluate the performance of AO numerically. In particular, two initialization strategies will be considered. A first simple initialization is based on the idea of equal-power precoding. For each user we select uniform power allocations \( x_k = \bar{x}_k \) and \( p_k = \bar{p}_k \) consuming the entire power budget \( P_{\text{max}} \), and satisfying the majorization constraint \( \log \bar{x}_k \preceq \log \bar{p}_k + \log \lambda_{Tk} \) in (6.14). This results in \( p_{k,n} = P_{\text{max}}/n_{Lk} \) for \( n = 1, \ldots, n_{Lk} \), and \( \bar{x}_k = \bar{p}_k \cdot \prod_n \lambda_{Tk,n}^{1/n_{Lk}} \). The second initialization strategy is adapted to running multiple instances of AO, for which it is desirable to generate a sample of \( x_k \)-vectors representing varying degrees of component spread as well as different component orders. We address the latter by letting \( \tilde{x}_k \) be a randomly selected permutation of the decreasing vector \( \bar{x}_{Lk} = \bar{p}_k \odot \lambda_{Tk} \). Regarding component spread, an initialization of
\( x_k \) is then composed as a convex mixture of the minimal-spread \( \bar{x}_k \) and the maximal-spread \( \tilde{x}_k \). In order to ensure that the resulting initialization satisfies \( \log x_k \preceq \log \bar{p}_k + \log \lambda_k \), we take the convex mixture in log-scale as \( x_k = e^{t \log \bar{x}_k + (1-t) \log \tilde{x}_k} \) for a random \( t \) uniformly distributed on \([0,1]\).

### 6.6 Numerical Results

This section evaluates numerically the proposed methods for joint precoder–decoding-order optimization. First, we investigate the performance gain of joint optimization compared to methods that omit precoder and/or decoding-order optimization. In addition, we put our fixed-ordering schemes to the test by comparing with different V-BLAST schemes that frequently update the decoding order based on short-term channel information [WFGV98, KVB05]. Second, the convergence of AO for single-mode optimization is analyzed. Throughout this section, we consider uplink transmission from four users (\( K = 4 \)), with either two or four antennas each, to an eight-antenna BS (\( n_R = 8 \)). The user power budgets (including path loss) are assumed to be ordered as \( P_{\text{max}1} \geq \ldots \geq P_{\text{max}K} \) and being logarithmically equispaced. Any such power-gain vector \([P_{\text{max}1} \ldots P_{\text{max}K}]^T\) is completely specified by its geometric-mean user power \( P_{\text{mean}} = \prod_{k=1}^K P_{\text{max}k}^{1/K} \) and user power spread \( \text{Pspread} = P_{\text{max}1}/P_{\text{max}K} \). This definition is related to the user distributions considered in [JB07].

We use the transmit-correlated Rayleigh channel model, with transmit correlation matrix \( \mathbf{R}_{Tk} \) based on a single parameter \( \rho_k \in [0,1] \) as \( [\mathbf{R}_{Tk}]_{mn} = \rho_k^{n-m} \) according to the exponential correlation model. We focus on four different performance measures: minimum user sum-rate (mRate) in (6.9); joint symbol error probability (JEP), the complementary probability of (6.10); maximum user weighted-sum MSE (mMSE) \( \max_k -u_k^{-\text{wMSE}} \), with \( u_k^{-\text{wMSE}} \) in (6.5); sum user weighted-sum MSE (sMSE) \( \sum_k -u_k^{-\text{wMSE}} \). Rate-normalizing weights \( \alpha_{k,n}^{\text{RN}} = 2^{R_{k,n}} - 1 \) are used with MSE-based functions in (6.5), where \( M\)-QAM signalling corresponds to a rate \( R = \log_2(M) \). In general, precoding problems on the form (6.6) are addressed as in [JOJ10] by reformulating (6.14) as a convex optimization problem. Ordering problems in Section 6.5.1 are addressed similarly as in [BDM09, Ch. 4.1.1] and solved using [glp12].

#### 6.6.1 Performance Analysis

We evaluate the performance of the proposed schemes with respect to four different objectives in Figures 6.4, 6.5, 6.6, and 6.7 using Monte Carlo simulation based on \( 10^5 \) channel realizations. We consider a 4-user uplink with 2-antenna users, and transmit correlation coefficients \( \rho_1 = 0.9, \rho_2 = 0.8, \rho_3 = 0.65, \rho_4 = 0.5 \). We use the following abbreviations:
6.6. NUMERICAL RESULTS

Figure 6.4: $P_{\text{mean}} = 10$ dB. $n_{L1} = \ldots = n_{L4} = 2$.

- “PM Opt.” and “MM Opt.”: optimal pure and mixed modes according to Section 6.4.
- “PM AO”: pure-mode optimization using AO with ten random initializations (Section 6.5.2).
- “Order 1” and “Order 2”: ZF-DF without precoding for decoding orders $\pi^1 = [1 \ldots n_L]^T$ and $\pi^2 = [n_L \ldots 1]^T$.
- “Prec. order 1” and “Prec. order 2”: counterparts of the former with optimal precoding in (6.6).
- “V-BLAST” and “RN V-BLAST”: ZF V-BLAST with ordering [WFGV98] and its rate-normalized equivalent [KVB05].
- “P.RN V-BLAST”: rate-normalized V-BLAST combined with optimal precoders of “PM Opt.”.

Figure 6.4 focuses on the user with minimum sum rate over its $n_{Lk} = 2$ subchannels. The two fixed decoding orders show similar performance when the users are close to equidistant. With increasing user spread, “Order 1” is clearly a favorable choice compared to “Order 2”, indicating that it is better to decode users further away later. This is natural, since being decoded later corresponds to an increased benefit of successive interference cancellation. Precoding has observably a positive effect on performance. However, comparing with “PM Opt.” reveals that even with precoding, both fixed orders are far from optimal up to a 10 dB spread. AO performs
close to optimally, and the optimal mixed mode performs only slightly better. With user spread exceeding 10 dB, it is observed that the user furthest away becomes the minimum-rate user, although most ordering gain is allocated to this user. The performance of the dynamically ordering V-BLAST scheme is clearly inferior to the optimized fixed-order designs. This is explained by the fact that V-BLAST is designed to provide short-term fairness among subchannels (by maximizing the minimum SNR), whereas the objective here is to provide long-term fairness on a user level.

In Figure 6.5 we consider the joint symbol error probability as performance measure. As argued in Section 6.4, mixed-mode optimization is unnecessary in this case since there is an optimal pure mode. With equidistant users (0 dB spread), we observe that a reversed order combined with precoding performs optimally among the fixed-order schemes: It is best to allocate most ordering gain to user 1 with the highest data rate of 8 bits/transmission. At the other extreme of 30 dB spread, most ordering gain should be allocated to user 4, which is furthest away, although this user only sends 2 bits/transmission. For intermediate power spreads, there is a large performance gain associated with optimizing both precoders and the decoding order. The plain V-BLAST scheme does not perform well in comparison, since it is not designed for the use of different constellations on the subchannels. This is successfully compensated for using the rate-normalized ordering in “RN V-BLAST”. Still, using an adaptive decoding order is clearly not as effective as optimal long-term precoding and
6.6. NUMERICAL RESULTS

Figure 6.6: $P_{\text{mean}} = 15$ dB. Users 1–4 transmit 8, 4, 4, 2 bits: $n_{L1} = 2$ (16-QAM), $n_{L2} = n_{L3} = 2$ (4-QAM), $n_{L4} = 1$ (4-QAM).

decoding ordering. Best performance is attained with “$P.RN\ V-BLAST$”; In this case, precoders optimized for the best fixed order are successfully combined with dynamic ordering. This is an interesting observation since the optimal precoding strategy in this case is unknown.

For the MSE-based utility functions we use the same transmission scenario as for the JEP in Figure 6.5. The evaluated schemes show similar behavior in Figures 6.6 and 6.7 as in Figures 6.4 and 6.5, respectively. Considering user fairness in terms of the maximum user MSE in Figure 6.6, it is worth mentioning that all dynamically ordering V-BLAST schemes fail to outperform the optimal fixed-ordering scheme, similarly as in Figure 6.4. Also, the additional gain of multi-mode transmission is again observed to be negligible.

6.6.2 Convergence of Alternating Optimization

We shall also evaluate the optimality and convergence of AO. We consider a 4-user uplink with 4-antenna users, based on $10^3$ test scenarios. Two different initializations: equal-power (EP) and random initialization (RI) using ten different samples. Performance objectives and randomly generated test scenarios are similar to those in Section 6.6.1 (in particular, the number of data streams are the same). For each user, a transmit-correlation parameter $\rho_k$ is drawn using a triangular PDF on the interval $[0, 1]$ in favor of high correlation. The power budgets $P_{\text{max1}} \geq \ldots \geq P_{\text{maxK}}$ are generated by randomly selecting $P_{\text{mean}}$ and $P_{\text{spread}}$. In the four cases studied, $P_{\text{mean}}$
is uniformly distributed (in dB) on the interval 5–15 dB. $P_{\text{spread}}$ is uniform on 0–20 dB for the minimum user data rate objective (“$m\text{Rate}$”), and uniform on 10–20 dB for the other three objectives. For each test scenario, the optimal pure-mode utility $\hat{U}_{PM}$ is computed, as well as maximized utilities $\hat{U}_{AO-EP}$ and $\hat{U}_{AO-RI}$ using AO with equal-power and random initialization, respectively. For both methods we compute the relative performance loss $R_\Delta = (\hat{U}_{PM} - \hat{U}_{AO})/|\hat{U}_{PM}|$.

Figure 6.8 displays the relative performance loss of AO, and provides several insights. Performance close to global optimality is more easily obtained for the two fairness-based objectives, which may have multiple global optima. Although global optimality is not always achieved, AO appears to perform very well in general. For three out of four objectives, it is unlikely that the relative difference exceeds a few percent after convergence. Higher values are observed for the JEP, which is natural. For example, a relative difference of 1 corresponds to twice the error probability, which is clearly not as severe as, e.g., losing the entire data rate, which would be the interpretation for the “$m\text{Rate}$” objective. It is also evident from Figure 6.8 that using multiple random initializations improves performance significantly.

Figure 6.9 displays the number of iterations $N_{\text{iter}}$ required for AO to converge (AO stops if the relative utility increase is less than $10^{-6}$). We conclude that AO in general only requires a handful of iterations to converge, whether based on equal-power initialization or random initialization (the ten-fold increase of the latter is due to the fact that the total number of iterations over the ten instances is shown).
6.6. NUMERICAL RESULTS

Figure 6.8: Relative performance loss of AO (compl. empirical CDF).

Figure 6.9: Number of iterations for AO (compl. empirical CDF).
6.7 Summary

We have investigated the joint precoder–decoding-order optimization problem in the multi-user MIMO uplink based on long-term channel information and general utility functions. The examination of ZF-DF equalization under channel fading, as described by a general non–line-of-sight MIMO channel model, resulted in two observations of practical interest for performance optimization. Firstly, the choice of precoder for a particular user does not statistically affect the other users’ performance. Secondly, the choice of decoding order does couple the users’ performance, since the various degrees of interference cancellation comprise a resource to be allocated to the users’ subchannels. Understanding the implications of these observations enabled straightforward system optimization both with and without time-sharing. Addressing the fact that such approaches are computationally demanding, we proposed a suboptimal method based on alternating optimization between precoding and ordering, and showed how to efficiently solve the ordering problem under transmit-correlated Rayleigh fading. Simulations demonstrated that alternating optimization could be effective in this context. Also, only a marginal performance increase of time-sharing could be observed. Fixed-ordering schemes were shown to perform better than common dynamically-ordering schemes when user fairness was in focus, while the reverse was true for other objectives of sum-performance type.
6. A Collection of Proofs

6.A.1 Proof of Theorem 6.1

To prove Theorem 6.1, we need the following result.

**Lemma 6.2.** Consider the block-diagonal matrix $T = \text{diag}(T_1, \ldots, T_K) \in \mathbb{C}^{M \times N}$, where $T_k \in \mathbb{C}^{M_k \times N_k}$ and $M_k \geq N_k$ for $k = 1, \ldots, K$. Let $\Pi = [\Pi_1 \ldots \Pi_K] \in \mathbb{R}^{N \times N}$ be a permutation matrix, with $\Pi_k \in \mathbb{R}^{N \times N_k}$, and $\Pi^{(k)} \in \mathbb{R}^{N_k \times N_k}$ is obtained by removing any row containing only zeros from $\Pi_k$. Let further $Q_{\Pi^{(k)}} L_{\Pi^{(k)}} = T_k$ be a $\Pi^{(k)}$-ordered QL decomposition of $T_k$ according to Def. 6.1. Form $Q = \text{diag}(Q_{\Pi^{(1)}}, \ldots, Q_{\Pi^{(K)}})$ and $L = \text{diag}(L_{\Pi^{(1)}}, \ldots, L_{\Pi^{(K)}})$. Then $QL = T$ is a $\Pi$-ordered QL decomposition of $T$.

**Proof.** It is evident that $QL = T$ and that $Q$ has orthonormal columns. It remains to verify that $\Pi L \Pi^T$ is lower triangular. This matrix can be expressed as

$$\Pi L \Pi^T = \sum_{k=1}^{K} \Pi_k L_{\Pi^{(k)}} \Pi_k^T = \sum_{k=1}^{K} I^k L^k I^k^T$$

if we define $I^k = \Pi_k \Pi^{(k)}^T$ and $L^k = \Pi^{(k)} L_{\Pi^{(k)}} \Pi^{(k)}^T$, the latter being lower triangular by definition. Note that $I^k$ is formed by removing $N - N_k$ columns from the $N \times N$ identity matrix by the definition of $\Pi^{(k)}$. Consider with $i < j$ the strictly upper-triangular element

$$\left[I^k L^k I^k^T\right]_{ij} = \sum_{m,n=1}^{N_k} \left[I^k\right]_{im} \left[I^k\right]_{jn} \left[L^k\right]_{mn}.$$  \hfill (6.23)

The matrix $L^k$ is by definition lower triangular rendering $[L^k]_{mn} = 0$ if $m < n$. For $m \geq n$, $I^k$ satisfies $\left[I^k\right]_{im} \left[I^k\right]_{jn} = 0$ if $i < j$. Thus, every term in the sum in (6.23) vanishes. We conclude that each $I^k L^k I^k^T$ is lower triangular, and the same holds for $\Pi L \Pi^T$. \hfill $\square$

We also need another result, which is a mere consequence of Definition 6.1.

**Lemma 6.3.** Let $QL = T$ be a $\Pi$-ordered QL decomposition of $T \in \mathbb{C}^{M \times N}$, where $M \geq N$. Then for any positive semi-definite $R \in \mathbb{C}^{M \times M}$, $l_{\Pi}^2(T^H R T) = l_{\Pi}^2(Q^H R Q) \odot l_{\Pi}^2(T^H T)$.

**Proof.** The proof is outlined as follows. Consider the two $\Pi$-ordered QL decompositions $QL = T$ and $Q_{RQ} L_{RQ} = R^{1/2} Q$. Setting $L_{RT} = L_{RQ} L$ it follows that $Q_{RQ} L_{RT}$ is a $\Pi$-ordered QL decomposition of $R^{1/2} T$.

$R^{1/2}QL = Q_{RQ}L_{RQ}L = Q_{RQ}L_{RT}$. The results follows from the fact that the diagonal of $L_{RT}$ is simply the Schur product of the diagonals of $L$ and $L_{RQ}$.

Consider first the $\Pi^{(k)}$-ordered QL decompositions $Q_{\Pi^{(k)}}L_{\Pi^{(k)}} = R_{Tk}^{1/2}P_k$ for $k = 1,\ldots,K$. Then by Lemma 6.2, the block-diagonal matrices $Q = \text{diag} (Q_{\Pi^{(1)}},\ldots,Q_{\Pi^{(K)}})$ and $L = \text{diag} (L_{\Pi^{(1)}},\ldots,L_{\Pi^{(K)}})$ comprise a $\Pi$-ordered QL decomposition of $R_{Tk}^{1/2}P$, where $R_T = \text{diag} (R_{T1},\ldots,R_{TK})$ and $P = \text{diag} (P_1,\ldots,P_K)$. It follows by Def. 6.1 that $x = l^2_{\Pi}(P_P^H R_{Tk} P_k)$ is given by $x = [x_1^T \ldots x_K^T]^T$, where $x_k = l^2_{\Pi^{(k)}}(P_k^H R_{Tk} P_k)$ for $k = 1,\ldots,K$.

Defining $\hat{H} = [\hat{H}_1 \ldots \hat{H}_K]$ and noting that $H = \hat{H}R_{Tk}^{1/2}$, Lemma 6.3 states that the SNR vector in (6.2) can be written as $\gamma = l^2_{\Pi}(Q_H^H \hat{H}^H \hat{H}Q) \odot x$. Since $Q$ is block-diagonal, we may write

$$\hat{H}Q = [\hat{H}_1 Q_{\Pi^{(1)}} \ldots \hat{H}_K Q_{\Pi^{(K)}}].$$

If $\hat{H}_k$ is unitarily invariant from the right (as under the general channel model), it follows that $\hat{H}_k Q_{\Pi^{(k)}}$ is equivalently distributed as the submatrix $H_{-k}^-$ of $\hat{H}_k$ consisting of its $n_{lk}$ first columns. With $\hat{H}_- = [\hat{H}_{-1} \ldots \hat{H}_{-K}]$, and statistically independent $\hat{H}_1,\ldots,\hat{H}_K$, it follows that $\gamma \sim l^2_{\Pi}(\hat{H}_-^H \hat{H}_-) \odot x$. With $\gamma = [\gamma_1 \ldots \gamma_K]^T$, the marginal SNR distribution for user $k$ depends on $P_k$ through $x_k$, and on $\pi$ through $x_k$ and $l^2_{\Pi}(\hat{H}_-^H \hat{H}_-)$.}

### 6.4.2 Proof of Theorem 6.2

The proof requires the utility region for a fixed decoding order, which is obtained in the following lemma.

**Lemma 6.4.** The utility region over all feasible pure communication modes with decoding order $\pi = [\pi_1^T \ldots \pi_K^T]^T$ is the hyper-rectangle in (6.7), provided that $u$ (M) is a normal utility function.

**Proof.** For a normal utility function $u$ (M), the precoder $P_k$ only affects the utility $u_k$ of user $k$, and it hence suffices to show for each user $k$ that the set of achievable utilities is given by the closed interval $\hat{u}_k (\pi) \leq u_k \leq \hat{u}_k (\pi)$. Denote the set of feasible precoders for user $k$ by $\mathcal{P}_k = \{ P_k \in \mathbb{C}^{n_{Tk} \times n_{lk}} : \text{tr} (P_k^H P_k) \leq P_{\text{max}k} \}$. By the continuity of the Cholesky decomposition it can be shown that $x_k = l^2_{\Pi^{(k)}}(P_k^H R_{Tk} P_k)$ defines a continuous mapping $x_k = x_k(P_k)$ on $P_k \in \mathcal{P}_k$ for any fixed $\pi$. Since $u_k(x_k,\pi)$ is continuous in $x_k$, and defined on the entire image set $x_k(\mathcal{P}_k)$, it follows that the composite mapping $u_k(x_k(P_k),\pi)$ is continuous in $P_k$ on $\mathcal{P}_k$. The set $\mathcal{P}_k$ is a compact and connected set, and the continuity
of \( u_k(x_k(P_k), \pi) \) ensures that these two properties hold for the image set \( u_k(x_k(P_k), \pi) \) as well. Hence, the set of achievable utilities for user \( k \) is necessarily a closed interval. That the interval is closed from above ensures that there exists a maximizer of (6.6) with corresponding maximal utility \( \hat{u}_k(\pi) \). The lower bound is \( \hat{u}_k(\pi) = u_k(0, \pi) \) due to the monotonicity of \( u_k(x_k, \pi) \) in \( x_k \), and is realized with \( P_k = 0 \). \( \square \)

Theorem 6.2 is a mere consequence of Lemma 6.4. Let \( M^* \) be a feasible pure communication mode, and denote its decoding order by \( \pi^* \). By Lemma 6.4, \( u(M^*) \in U_{\pi^*}^{PM} \) and it follows that \( u(M^*) \in \bigcup_{\pi \in \mathcal{S}_*} U_{\pi}^{PM} \). Reversely, let \( \hat{u} \in \bigcup_{\pi \in \mathcal{S}_*} U_{\pi}^{PM} \) be a given utility vector. Then \( \hat{u} \in U_{\pi}^{PM} \) for some decoding order \( \pi \). Lemma 6.4 ensures the existence of a pure mode \( \hat{M} \) with decoding order \( \hat{\pi} \) and utility \( u(\hat{M}) = \hat{u} \).

6.A.3 Proof of Theorem 6.3

Let \( \mathcal{M}^* \) be a feasible mixed communication mode with \( \Pr (\mathcal{M} = M^*_i) = \theta_i^* \) for \( i = 1, \ldots, I \), and \( \sum_{i=1}^I \theta_i^* = 1 \). Then \( u_{\mathcal{M}^*} = \mathbb{E}[u(M^*)] = \sum_{i=1}^I \theta_i^* u(M^*_i) \). By Theorem 6.2, \( u(M^*_i) \in U_{\pi}^{PM} \) for each \( i = 1, \ldots, I \). Since \( u_{\mathcal{M}^*} \) is a convex mixture of elements in \( U_{\pi}^{PM} \), it follows that \( u_{\mathcal{M}^*} \in \text{conv}(U_{\pi}^{PM}) \). Reversely, let \( \hat{u} \in \text{conv}(U_{\pi}^{PM}) \) be a given utility vector. By Carathéodory’s theorem, \( \hat{u} \) can be written as a convex mixture of \( I \leq K + 1 \) points in \( U_{\pi}^{PM} \). Hence, let \( \hat{u} = \sum_{i=1}^I \tilde{\theta}_i \hat{u}_i \) with \( \sum_{i=1}^I \tilde{\theta}_i = 1 \) and \( \hat{u}_i \in U_{\pi}^{PM} \) for \( i = 1, \ldots, I \). By Theorem 6.2 there are pure modes \( \mathcal{M}_1, \ldots, \mathcal{M}_I \) with utilities \( u(\mathcal{M}_i) = \hat{u}_i \). It follows that the mixed mode \( \mathcal{M} \) with \( \Pr (\mathcal{M} = \mathcal{M}_i) = \tilde{\theta}_i \) for \( i = 1, \ldots, I \) has \( u_{\mathcal{M}} = \hat{u} \).

6.A.4 Proof of Lemma 6.1

Let \( u^* \in U_{\pi}^{PM} \). Since \( U_{\pi}^{PM} = \bigcup_{\pi \in \mathcal{S}_*} U_{\pi}^{PM} \), there is a decoding order \( \pi^* \) such that \( u^* \in U_{\pi^*}^{PM} \). By Lemma 6.4, \( u^* \leq \hat{u}(\pi^*) \) component-wise and the result follows since \( \hat{u}(\pi^*) \in \mathcal{V} \). Similarly, let \( \hat{u} \in U_{\pi}^{MM} \). By Carathéodory’s theorem, let \( \hat{u} = \sum_{i=1}^I \tilde{\theta}_i \hat{u}_i \) with \( \sum_{i=1}^I \tilde{\theta}_i = 1 \), and \( \tilde{\theta}_i > 0 \) and \( \hat{u}_i \in U_{\pi}^{PM} \) for \( i = 1, \ldots, I \). Then by Theorem 6.2 there are decoding orders \( \tilde{\pi}^1, \ldots, \tilde{\pi}^I \) such that \( \hat{u}_i = U_{\tilde{\pi}^i}^{PM} \) for \( i = 1, \ldots, I \). It follows by Lemma 6.4 that \( \hat{u} \leq \sum_{i=1}^I \tilde{\theta}_i \hat{u}(\tilde{\pi}^i) \in \text{conv}(\mathcal{V}) \).

6.A.5 Proof of Theorem 6.4

Let \( u^* \) denote a maximizer of \( \mathcal{H}(u) \) over \( u \in \mathcal{V} \) in (6.8). Since \( u^* \in U_{\pi}^{PM} \), Theorem 6.2 states that there exists a pure mode \( M^* \) having \( u(M^*) = u^* \). Next, let \( \mathcal{M} \) be an arbitrary feasible mixed mode with possible outcomes \( M_1, \ldots, M_I \). Applying Jensen’s inequality results in \( \mathcal{H}(\mathbb{E}[u(M)]) \leq \mathbb{E}[\mathcal{H}(u(M))] \) when \( \mathcal{H}(u) \) is convex. Thus it suffices to prove the result for
$U_M = \mathbb{E}[\mathcal{H}(u(M))]$ for a general $\mathcal{H}(u)$. Let $i_{\text{max}} = \text{arg max}_i \mathcal{H}(u(M_i))$, and note that $\mathbb{E}[\mathcal{H}(u(M))] \leq \mathcal{H}(u(M_{i_{\text{max}}}))$. By Lemma 6.1, there is a $\tilde{u} \in \mathcal{V}$ such that $u(M_{i_{\text{max}}}) \leq \tilde{u}$. By the monotonicity of $\mathcal{H}(u)$, $\mathcal{H}(u(M_{i_{\text{max}}})) \leq \mathcal{H}(\tilde{u})$. Lastly, we note that $\mathcal{H}(\tilde{u}) \leq \mathcal{H}(u^*)$ by the definition of $u^*$. This concludes $U_{M^*} \geq U_M$ in both cases.
Chapter 7

Conclusions and Future Work

7.1 Conclusions

This thesis revolved around the question of how to perform MIMO transceiver design under practical assumptions in wireless communications; the time-varying nature of wireless channels does not harmonize with assuming perfect channel information at the transmitter. Having instead a statistical description of small-scale fading available at the transmitter makes the associated design problem inherently more complex. One conclusion of the thesis is that zero-forcing DF transceivers are amenable to joint transmitter-receiver optimization in this context, in the sense that tractable optimization problems can be posed.

Spatial correlation of wireless MIMO channels, resulting from insufficient antenna spacing, was taken into account for transceiver design. From an optimization point of view, allowing for spatial correlation on the transmitter side leads to similar challenges as those faced under a perfect-CSI assumption. The recent advances in the field of MIMO transceiver design relied on mathematical tools such as majorization theory and the generalized triangular decomposition, leading to convex optimization problems or even closed-form solutions. This thesis demonstrated how these tools can be applied to solve similar problems with only statistical transmitter CSI.

Allowing for additional spatial correlation on the receiver side, this thesis recognized that there is a gap in the literature on performance measures for the ZF-DF transceiver in correlated fading. To this end, a study was initiated that resulted in novel expressions based on Newton’s divided differences. Being a mature field, this connection provides means for analytical manipulations as well as numerically stable computation of such performance measures.

The thesis also dealt with the multi-user MIMO uplink, a setting that brings a new aspect for selecting a detection order for successive interfer-
enue cancellation: the benefit of interference cancellation can be regarded as a resource to be allocated to different users via the choice of detection order. The thesis posed the problem of jointly optimizing the users’ precoders as well as the detection order. The proposed solution was to employ alternating optimization between precoding and ordering. It was shown that jointly optimizing the users’ precoders could be reduced to solving single-user precoding problems. The combinatorial problem of optimizing the detection order was cast into the framework of linear assignment problems, enabling efficient algorithms to be employed.

It was also observed throughout the thesis that MSE-based performance measures lead to simple solutions to the optimization problems. In the single-user case, an exact solution is directly obtained without resorting to iterative algorithms of convex optimization. This is particularly interesting, since a straightforward solution to the transceiver design problem leaves room for tuning other parameters, such as the data-stream rate allocation, in conjunction with optimal precoding. This is also important for optimization in the multi-user uplink, where joint transceiver design relies on iteratively solving single-user precoding problems. Moreover, it was observed that the MSE-optimal detection order is trivial to find, and the overall complexity for MSE-based optimization in the multi-user MIMO uplink is negligible.

7.2 Future Work

There are a number of extensions to the material covered in the thesis that remain to be investigated. The following list contains a small selection:

- **Rician channel model**: Only zero-mean MIMO channel models are treated in this thesis, specifically the correlated Rayleigh and double-scattering models for multipath fading. Adopting a general Rician model instead would include a line-of-sight component as well. Moreover, a non-zero channel mean may also represent imperfect CSI in terms of a channel estimate at the transmitter. The statistical model may account for errors due to channel estimation, quantization and feedback delay in an FDD system, for example.

- **Optimal regime for spatial multiplexing**: This thesis considers MIMO transceivers for spatial multiplexing (single-stream beamforming is also included). However, the spatial-multiplexing capabilities of a wireless MIMO channel reduce as the degree of spatial correlation increases. It would be interesting to characterize when the precoder-optimized MIMO transceivers based on spatial multiplexing should be replaced by, for example, space-time block coding.
7.2. FUTURE WORK

- *Joint optimization of precoder and rate allocation:* It was observed numerically in Chapter 3 that there is a large performance gain associated with jointly optimizing the precoder and the data-stream rate allocation. This thesis showed how to optimize the precoder for a given rate allocation, but efficient methods for joint optimization have not been considered. Since our publication in [JOJ10] such an approach has been initiated in [LP12].

- *Performance measures for ZF-DF in the multi-user MIMO uplink:* Novel expressions for evaluating the performance of the single-user ZF-DF transceiver were presented in Chapter 4 for spatially correlated MIMO channels. The approach leading to these results is still in its infancy, and we believe that corresponding expressions can be found for the multi-user MIMO uplink.
# Nomenclature

## Abbreviations and Acronyms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AO</td>
<td>Alternating optimization</td>
</tr>
<tr>
<td>AP</td>
<td>Assignment problem</td>
</tr>
<tr>
<td>AWGN</td>
<td>Additive white Gaussian noise</td>
</tr>
<tr>
<td>BEP</td>
<td>Bit error probability</td>
</tr>
<tr>
<td>BLAST</td>
<td>Bell Labs layered space time</td>
</tr>
<tr>
<td>BS</td>
<td>Base station</td>
</tr>
<tr>
<td>CDMA</td>
<td>Code-division multiple access</td>
</tr>
<tr>
<td>CSI</td>
<td>Channel-state information</td>
</tr>
<tr>
<td>CSIT</td>
<td>CSI at the transmitter</td>
</tr>
<tr>
<td>dB</td>
<td>Decibel</td>
</tr>
<tr>
<td>DF</td>
<td>Decision feedback</td>
</tr>
<tr>
<td>DS</td>
<td>Double scattering</td>
</tr>
<tr>
<td>DSL</td>
<td>Digital subscriber line</td>
</tr>
<tr>
<td>FDD</td>
<td>Frequency-division duplexing</td>
</tr>
<tr>
<td>FDMA</td>
<td>Frequency-division multiple access</td>
</tr>
<tr>
<td>GTD</td>
<td>Generalized triangular decomposition</td>
</tr>
<tr>
<td>i.i.d.</td>
<td>Independent and identically distributed</td>
</tr>
<tr>
<td>JEP</td>
<td>Joint symbol error probability</td>
</tr>
<tr>
<td>KKT</td>
<td>Karush Kuhn Tucker</td>
</tr>
<tr>
<td>LBAP</td>
<td>Linear bottleneck assignment problem</td>
</tr>
<tr>
<td>LSAP</td>
<td>Linear sum assignment problem</td>
</tr>
<tr>
<td>LTE</td>
<td>Long Term Evolution</td>
</tr>
<tr>
<td>MF</td>
<td>Matched filter</td>
</tr>
<tr>
<td>MGF</td>
<td>Moment generating function</td>
</tr>
<tr>
<td>MIMO</td>
<td>Multiple-input multiple-output</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum likelihood</td>
</tr>
<tr>
<td>MM</td>
<td>Mixed communication mode</td>
</tr>
<tr>
<td>MMSE</td>
<td>Minimum MSE</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean square error</td>
</tr>
<tr>
<td>OFDM</td>
<td>Orthogonal frequency-division multiplexing</td>
</tr>
</tbody>
</table>
OFDMA      Orthogonal frequency-division multiple access
PDF        Probability density function
PM         Pure communication mode
QAM        Quadrature amplitude modulation
QoS        Quality of service
QL         A matrix decomposition (Not an acronym)
SDMA       Space-division multiple access
SEP        Symbol error probability
SIC        Successive interference cancellation
SISO       Single-input single-output
SNR        Signal-to-noise ratio
STBC       Space-time block code
SVD        Singular-value decomposition
TDD        Time-division duplexing
TDMA       Time-division multiple access
V-BLAST     Vertical BLAST
ZF          Zero forcing
WLAN       Wireless local area network

Mathematical Notation

\( \mathbb{R} \)          The set of real numbers.
\( \mathbb{R}_+ \)         The set of non-negative real numbers.
\( \mathbb{C} \)          The set of complex numbers.
\( \mathbb{R}^N \)         The set of real \( N \)-vectors.
\( D_\downarrow \)        The set of decreasing vectors,
                          \( \{ x \in \mathbb{R}^N : x_1 \geq \ldots \geq x_N \} \).
\( D_\uparrow \)        The set of increasing vectors,
                          \( \{ x \in \mathbb{R}^N : x_1 \leq \ldots \leq x_N \} \).
\( \mathbb{C}^N \)         The set of complex \( N \)-vectors.
\( \mathbb{R}^{N \times M} \) The set of real \( N \times M \)-matrices.
\( \mathbb{C}^{N \times M} \) The set of complex \( N \times M \)-matrices.

\(|x|\)                  Absolute value of a scalar \( x \).
\( \lfloor x \rfloor \) The largest integer smaller than the scalar \( x \in \mathbb{R} \).

\( x_i = [x]_i \)       The \( i \)th element of a vector \( x \).
\( x_\downarrow \)       The vector \( x \) rearranged in decreasing order.
\( x_{\downarrow i} \) The \( i \)th largest component of the real vector \( x \).
\( x_\uparrow \)       The vector \( x \) rearranged in increasing order.
\( x_{\uparrow i} \) The \( i \)th smallest component of the real vector \( x \).
| NOMENCLATURE | 147 |

| **|| **x** | p** || The p-norm of a vector x (i.e., \(||x||_p = (\sum_i |x_i|^p)^{1/p}\)). |
| **|| **x** | 2** || The Euclidean norm \(||x||_2\). |
| **e**x** || The vector \((e^{x_1}, \ldots, e^{x_N})\). |
| **log(x)** || The vector \((\log(x_1), \ldots, \log(x_N))\). |
| **x \preceq y** || x is majorized by y (see [MOA11]). |
| **x \prec_w y** || x is weakly submajorized by y. |
| **x \prec_w y** || x is weakly supermajorized by y. |
| **x \odot y** || Componentwise (Schur) product between x and y. |
| **x \leq y** || Means that \(x_i \leq y_i\) for all vector indices i. |

| **x_{ij} = [X]_{ij}** || Two ways of writing element \((i, j)\) of a matrix \(X\). |
| **\text{diag} (x_1, \ldots, x_N)** || The diagonal matrix with \(x_1, \ldots, x_N\) on the diagonal. |
| **X^T** || The transpose of \(X\). |
| **X^H** || The Hermitian transpose of \(X\). |
| **X^{-1}** || The inverse of a square matrix \(X\). |
| **\text{det} (X)** || The determinant of a square matrix \(X\). |
| **\text{vec} (X)** || The vector obtained by stacking the columns of \(X\). |
| **\text{tr} (X)** || The trace of a square matrix \(X\). |
| **\text{span}(X)** || The linear space spanned by the columns of \(X\). |
| **\text{QL} = T** || The QL decomposition of \(T \in C^{M \times N}(M \geq N)\), where \(Q \in C^{M \times N}\) is semi-unitary \((Q^HQ = I)\), and \(L \in C^{N \times N}\) is lower triangular with real, non-negative diagonal entries. |

| **E [X]** || The mathematical expectation of a random \(X\). |
| **Pr (X)** || The probability of an event \(X\). |
| **CN (x, R)** || The circular-symmetric complex Gaussian distribution with mean \(x\) and covariance \(R\). |

| **\subset** || Subset. |
| **\in** || Belongs to. |
| **\forall x** || For all. |
| **|S||** || The cardinality (i.e., number of members) of a set \(S\). |

| **I** || The identity matrix. |
| **1** || The vector of only ones. |
| **0** || The vector of only zeros. |
Bibliography


BIBLIOGRAPHY


