Pontifícia Universidade Católica
do Rio de Janeiro

Erico de Souza Prado Lopes

Discrete Precoding and Adjusted Detection for Multiuser MIMO Systems with PSK Modulation

Dissertação de Mestrado

Dissertation presented to the Programa de Pós–graduação em Engenharia Elétrica of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia Elétrica.

Advisor: Prof. Lukas T. N. Landau

Rio de Janeiro
March 2021

Pontifícia Universidade Católica
do Rio de Janeiro

Erico de Souza Prado Lopes

Discrete Precoding and Adjusted Detection for Multiuser MIMO Systems with PSK Modulation

Dissertation presented to the Programa de Pós–graduação em Engenharia Elétrica of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia Elétrica. Approved by the Examination Committee.

Prof. Lukas T. N. Landau
Advisor
Centro de Estudos em Telecomunicações (CETUC) – PUC-Rio

Prof. Inbar Fijalkow
Université Cergy-Pontoise – UCP

Prof. Andreas Fischer
Technische Universität Dresden – TU Dresden

Prof. Amine Mezghani
University of Manitoba – UM

Prof. Rodrigo C. de Lamare
Centro de Estudos em Telecomunicações (CETUC) – PUC-Rio

Rio de Janeiro, March the 3rd, 2021

Erico de Souza Prado Lopes

Majored in Electrical Engineering by the Pontifical Catholic University of Rio de Janeiro (Rio de Janeiro, Brazil).

Bibliographic data

Lopes, Erico de Souza Prado

Discrete Precoding and Adjusted Detection for Multiuser MIMO Systems with PSK Modulation / Erico de Souza Prado Lopes; advisor: Lukas T. N. Landau. – 2021.

v., 74 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica.

Inclui bibliografia

1. Engenharia Elétrica – Teses. 2. Precodificação de Baixa Resolução. 3. Quantização em Fase. 4. Sistemas MIMO. 5. Método Branch-and-Bound. 6. Detecção e Decodificação Iterativa. I. Landau, Lukas T. N.. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. III. Título.

CDD: 621.3

To my father, mother and girlfriend for their support
and encouragement.

Acknowledgments

First, I would like to thank Prof. Lukas T. N. Landau for giving me this great opportunity and introducing me to this novel topic. During these two years, Prof. Lukas taught me a lot, not only from the academic world but also from life. This thesis would not be possible without his guidance, excitement and help. Moreover, I would like to thank him for letting me borrow his computer during the COVID-19 pandemic, this thesis would not have been possible without it.

I would like to thank Prof. Inbar Fijalkow, Prof. Andreas Fischer, Prof. Amine Mezghani, and Prof. Rodrigo de Lamare for serving as referees in the defense committee. I also would like to thank them for their contributions to this study and for the inspiring and productive comments during the defense. Finally, I would like to thank for their compliments, it meant a lot that such successful professors recognized my work.

I would like to thank my mother, father, and girlfriend for being the most wonderful people I know. Their support was what made me go on and find strength in difficult times. I would like to thank my mom for being the person that always lift up my spirit when I was down. I would like to thank my dad for his advice and for his patience when I didn't hear. At last, I would like to thank my girlfriend, Marianne, for the affection given in difficult times. This thesis is also yours, I love you guys.

I had the opportunity to share a laboratory with some amazing people. I would like to thank Ali, Emílio, Thiago, and Bia for a great time studying together and for their friendship. Thanks also to Diana, who was a funny and enjoyable desk neighbor.

I would like to thank Prof. Rodrigo C. de Lamare for his words of support during difficult times and for highlighting the importance of pursuing a master's degree when I was an undergraduate student.

My sincere thanks to all professors, students, and staff from CETUC for providing a pleasant environment for me to pursue my studies. Thanks to all of you CETUC became my second home in my four years of studying telecommunications.

Finally, I would like to thank CNPq for financial support to complete my master's degree. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Abstract

de Souza Prado Lopes, Erico; Landau, Lukas Tobias Nepomuk (Advisor). **Discrete Precoding and Adjusted Detection for Multiuser MIMO Systems with PSK Modulation** . Rio de Janeiro, 2021. 74p. Dissertação de mestrado – Departamento de Centro de Estudos em Telecomunicações (CETUC), Pontifícia Universidade Católica do Rio de Janeiro.

With an increasing number of antennas in multiple-input multiple-output (MIMO) systems, the energy consumption and costs of the corresponding front ends become relevant. In this context, a promising approach is the consideration of low-resolution data converters. In this study two novel optimal precoding branch-and-bound algorithms constrained to constant envelope signals and phase quantization are proposed. The first maximizes the minimum distance to the decision threshold (MMDDT) at the receivers, while the second minimizes the MSE between the users' data symbols and the receive signal. This MMDDT design presented in this study is a generalization of prior designs that rely on 1-bit quantization. Moreover, unlike the prior MMSE design that relies on 1-bit resolution, the proposed MMSE approach employs uniform phase quantization and the bounding step in the branch-and-bound method is different in terms of considering the most restrictive relaxation of the nonconvex problem, which is then utilized for a suboptimal design also. Moreover, three different soft detection methods and an iterative detection and decoding scheme that allow the utilization of channel coding in conjunction with low-resolution precoding are proposed. Besides an exact approach for computing the extrinsic information, two approximations with reduced computational complexity are devised. The proposed branch-and-bound precoding algorithms are superior to the existing methods in terms of bit error rate. Numerical results show that the proposed approaches have significantly lower complexity than exhaustive search. Finally, results based on an LDPC block code indicate that the proposed receive processing schemes yield a lower bit-error-rate compared to the conventional design.

Keywords

Precoding; Low-Resolution Quantization; Phase Quantization; MIMO systems; Branch-and-Bound methods; MMSE; MMDDT; Constant Envelope signals; Log-Likelihood-Ratios; Soft Detection; Iterative Detection and Decoding.

Resumo

de Souza Prado Lopes, Erico; Landau, Lukas Tobias Nepomuk. **Precodificação Discreta e Detecção Correspondente para Sistemas MIMO Multiusuário que utilizam Modulação PSK** . Rio de Janeiro, 2021. 74p. Dissertação de Mestrado – Departamento de Centro de Estudos em Telecomunicações (CETUC), Pontifícia Universidade Católica do Rio de Janeiro.

Com um número crescente de antenas em sistemas MIMO, o consumo de energia e os custos das interfaces de rádio correspondentes tornam-se relevantes. Nesse contexto, uma abordagem promissora é a utilização de conversores de dados de baixa resolução. Neste estudo, propomos dois novos pré-codificadores ótimos para a sinais de envelope constante e quantização de fase. O primeiro maximiza a distância mínima para o limite de decisão (MMDDT) nos receptores, enquanto o segundo minimiza o erro médio quadrático entre os símbolos dos usuários e o sinal de recepção. O design MMDDT apresetado nesse estudo é uma generalização de designs anteriores que baseiam-se em quantização de 1-bit. Além disso, ao contrário do projeto MMSE anterior que se baseia na resolução de 1-bit, a abordagem proposta emprega quantização de fase uniforme e a etapa de limite no método branch-and-bound é diferente em termos de considerar o relaxamento mais restritivo do problema não convexo, que é então utilizado para um design sub ótimo também. Além disso, três métodos diferentes de detecção suave e um esquema iterativo de detecção e decodificação que permite a utilização de codificação de canal em conjunto com pré-codificação de baixa resolução são propostos. Além de uma abordagem exata para calcular a informação extrínseca, duas aproximações com reduzida complexidade computacional são propostas. Os algoritmos propostos de pré-codificação branch-and-bound são superiores aos métodos existentes em termos de taxa de erro de bit. Resultados numéricos mostram que as abordagens propostas têm complexidade significativamente menor do que a busca exaustiva. Finalmente, os resultados baseados em um código de bloco LDPC indicam que os esquemas de processamento de recepção geram uma taxa de erro de bit menor em comparação com o projeto convencional.

Palavras-chave

Precodificação; Quantização de Baixa Resolução; Quantização em Fase; Sistemas MIMO; Método Branch-and-Bound; MMSE; MMDDT; Sinais de Envelope Constante; Log da Razão de Verossimilhança; Detecção Suave; Detecção e Decodificação Iterativa.

Table of contents

List of figures

List of tables

List of Abreviations

ADC – Analog-to-Digital Converter
AWGN – Additive White Gaussian Noise
BER – Bit Error Rate
BS – Base station
B&B – Branch-and-Bound
CE – Constant Envelope
CSI – Channel State Information
DAC – Digital-to-Analog Converter
DL – Downlink
DPA – Discrete Precoding Aware
DPA-LM – Discrete Precoding Aware Linear Model
EE – Energy Efficiency
GDPA – Gaussian Discrete Precoding Aware
IDD – Iterative Detection and Decoding
IPM – Interior Points Method
LDPC – Low Density Parity Check
LLR – Log-Likelihood-Ratio
LP – Linear Program
MAP – Maximum a posteriori Probability
MIMO – Multiple-Input Multiple-Output
MDDT – Minimum Distance to the Decision Threshold
MMDDT – Maximum of the Minimum Distance to the Decision Threshold
MSE – Mean Square Error
MMSE – Minimum Mean Square Error
MSM –Maximum Safety Margin
MUI – Multiuser Interference
PA – Power Amplifier
PAPC – Per Antenna Power Constraint
PDF – Probability Density Function
PSK – Phase Shift Keying
QAM – Quadrature Amplitude Modulation
QPSK – Quadrature Phase Shift Keying

RFFE – Radio Frequency Front End
SIC – Successive Interference Cancellation
SISO – Single-Input Single-Output
SNR – Signal-to-noise ratio
SOCP – Second Order Cone Program
TPC – Total Power Constraint
WFQ – Wiener Filter Quantized
ZF – Zero Forcing
ZF-P – Phase Zero Forcing

Published and Submitted Articles

The work presented in this thesis gave rise to the following papers:

1. *Optimal Precoding for Multiuser MIMO Systems With Phase Quantization and PSK Modulation via Branch-and-Bound* published in IEEE Wireless Communications Letters, in September 2020.
2. *Optimal and Suboptimal MMSE Precoding for Multiuser MIMO Systems Using Constant Envelope Signals with Phase Quantization at the Transmitter and PSK Modulation* published in the proceedings of the 24th International ITG Workshop on Smart Antennas (WSA 2020).
3. *Iterative Detection and Decoding for Multiuser MIMO Systems with Low Resolution Precoding and PSK Modulation* published in the proceedings of IEEE Statistical Signal Processing Workshop 2021 (SSP 2021).
4. *Discrete MMSE Precoding for Multiuser MIMO Systems with PSK Modulation* submitted to IEEE Transactions on Wireless Communications in the 29th, December 2021.

1
Introduction

Multiuser multiple-input multiple-output (MIMO) systems have become central in the wireless communications area. Many standards, such as IEEE 802.11n (Wi-Fi), IEEE 802.11ac (Wi-Fi), HSPA+ (3G), WiMAX, and Long Term Evolution (4G LTE), have adopted MIMO due to its benefits over single-input single-output (SISO) systems.

MIMO systems have considerable advantages over SISO, e.g. increased capacity, reliability and spectrum efficiency [1]. Thus, multiuser MIMO systems are expected to be vital for the future of wireless communications [2].

1.1
Motivation and Context

Although MIMO is considered as essential for the future of wireless communications, the energy consumption and costs related to having multiple radio frequency front ends (RFFEs) present a challenge for this kind of technology [3].

The power consumption of one individual RFFE generally does not generate a huge impact on the total energy consumption of the system, yet when a large number of RFFEs is employed its energy consumption can become significant and can cause energy efficiency (EE) problems.

Energy efficiency (EE) is a key requirement for the next generation of wireless communications. According to [4], 6G networks will require 10 to 100 times higher EE when compared to 5G, to enable scalable low-cost deployments, with low environmental impact, and better coverage. As stated in [5, 6], another central demand for future networks is higher data reliability.

With this, a challenge for the design of MIMO systems is the reduction of the energy consumption and costs related to the large number of RFFEs with minimum bit-error-rate (BER) compromise.

1.1.1
Radio Frequency Front End Components versus Energy Efficiency

To realize low energy consumption and low hardware-related costs different approaches have risen in literature for increasing the EE of the most

consuming elements of a RFFE. The following will give a brief overview on these elements and the main solutions for the increasing of their EE.

1.1.1.1 Power Amplifier

The power amplifier (PA) in wireless communications systems is the device responsible to amplify the signal and drive it to the antenna. Ideally the PA should transfer all the energy it consumes to the amplification of the arriving signal, providing to the antenna an amplified copy of the its input. However, in real devices a trade-off between distortion and EE has to be made. For this reason, the PAs are characterized by two main features, linearity and efficiency.

PAs, when operating with minimum distortion, generally have relatively low EE, e.g. a class A amplifier transfers a maximum of 50% of the consumed power to its output [7]. This problem is enhanced when high peak-to-average wave forms are considered [8].

Thus, one approach for the increasing the PA's EE in the consideration of constant envelope (CE) waveforms, such as the ones used for PSK modulation.

1.1.1.2 Digital-to-Analog Converter

On the other side, depending on the pathloss, the converters can be one of the most energy consuming elements of a RFFE. A noteworthy fact regarding the consumption of digital-to-analog converters (DACs) is that it scales exponentially with its resolution in amplitude [9].

In the downlink (DL), the baseband signal is generated by two digital-to-analog converters (DACs). Traditional RFFEs are generally employ to a pair of high resolution DACs (e.g. 10 bits or more). The mentioned behavior between the DAC's resolution in amplitude and its energy consumption and the fact that massive multiuser MIMO systems generally have hundreds or thousands of BS antennas yield a prohibitive energy consumption when employing high-resolution converters [10]. Hence, adopting low resolution DACs is a potential solution for mitigating the high energy consumption issue.

However, the adoption of low-resolution converters can cause significant performance degradation in the BER. In this context, a promising method to balance this BER loss and approach the requirements of future wireless communications networks [5, 6] is the usage of channel coding.

1.2 Related Work

Another way to mitigate the performance degradation caused by coarse quantization is the employment of low-resolution precoding and detection schemes. These techniques have been receiving increasing attention from the wireless communications community.

Several precoding strategies with low-resolution data converters exist in literature. Linear approaches, such as the phase Zero-Forcing (ZF-P) precoder [11] and the Wiener Filter Quantized (WFQ) precoder [12], benefit from a relatively low computational complexity. However, they yield performance degradation in BER especially for higher-order modulation [13, 14, 15, 16]. Therefore, more sophisticated nonlinear approaches have been presented recently in [17, 18, 19, 20, 21, 22, 23, 24]. However, the methods from [17, 18, 19, 20, 21, 22, 23] imply rounding and the method from [24] implies the convergence to a local minimum, which make the solution provided by these approaches suboptimal in their design criteria.

Moreover, some optimal precoders exist in literature. In [25] a branch-and-bound algorithm was developed for maximizing the minimum distance to the decision threshold (MMDDT) for the 1-bit case. In addition, in [26] a branch-and-bound algorithm, is presented for finding the transmit vector that minimizes the mean square error (MMSE), also for the 1-bit case.

On the detection side, various different techniques with low-resolution analog-to-digital converters (ADCs) were developed, e.g. [27, 28, 29, 30, 31, 32, 33].

1.3 Contributions

The scope of this work lies in the design of precoding and detection schemes for a MIMO downlink (DL) system in which the base station (BS) is considered to use PSK modulation in conjunction with uniform phase quantizers that have an arbitrary number of quantization regions. With this in mind this section is divided into two parts, the first exposes the contributions made in the precoding side, while the second explains the ones made in the detection side.

1.3.1 Precoding

This work devises an optimal precoding strategy based on the MMDDT criterion and an optimal and a suboptimal precoding techniques based on the

MMSE criterion.

1.3.1.1 MMDDT Contributions

In the present study the work from [25], which uses 1-bit quantization, is generalized for phase quantizers with arbitrary number of phases at the transmit antennas and PSK modulation. This extension should be considered as non trivial because in the case of PSK, each symbol cannot be decomposed in independent real and imaginary part as done in the 1-bit case. The proposed precoder is optimal in terms of the MMDDT criterion, obtained by using a sophisticated branch-and-bound strategy.

In contrast to the method presented in [20], which implies a rounding step, the proposed algorithm finds the global optimum. Rounding steps when used for computing the final precoding vector causes BER performance degradation and can generate error floors. Moreover, unlike the problem formulation in [20] with 2^q different phases, the proposed approach supports an arbitrary number of phase quantizations. In the initial step of the proposed method the relaxed problem is solved and rounded to the feasible set. Subsequently the optimum is determined by a tree search based algorithm. The MMDDT ow-resolution precoding contributions were published in [34].

1.3.1.2 MMSE Contributions

In this study, a novel branch-and-bound algorithm for finding the vector that yields the minimum MSE is also devised. In contrast to the MMDDT criterion considered in [20] and [25], which is promising in the context of hard decision receivers and the high signal-to-noise ratio (SNR) regime, the MMSE criterion is more general.

Besides the consideration of phase quantization and PSK modulation, the proposed approach uses a different bounding method as the approach presented in [26]. Whereas the approach in [26] employs for relaxation the constant envelope constraint, the present study relies on the convex hull of the discrete non-convex feasible set, which is by definition the most restrictive relaxation for establishing convexity. Consequently, the lower bounds computed in the present study are greater or equal to the ones computed in [26].

In addition, a suboptimal precoding approach is devised based on the relaxed problem, which is formulated as a convex quadratic program. Note that this method is different from the one presented in [19], since, in the present study, the scaling factor is also part of the optimization problem and, thus,

does not need to be approximated as done in [19]. Furthermore, while the method from [19] implies QPSK data modulation and 2^q quantization phases, the proposed approach allows the entries of the data vector to belong to any kind of PSK modulation and the quantizers to have an arbitrary number of phases. The MMSE low-resolution precoding contributions were published in [35, 36].

The numerical results show that the MMSE branch-and-bound method corresponds to a lower uncoded BER in comparison to the state-of-the-art algorithms for the low and intermediate SNR regime. Moreover, the numerical results confirm that when operating in low SNR, only a small number of bounds need to be evaluated to determine the optimal solution.

1.3.2 Detection

For receiving, three novel discrete precoding aware (DPA) soft detection methods integrated in an iterative detection and decoding (IDD) scheme are proposed. To the best of our knowledge, the present study is the first which considers soft detection for a MIMO DL system with discrete precoding.

The three proposed soft detection approaches calculate extrinsic information values, which are used for computing the log-likelihood-ratios (LLR) via the DPA-IDD algorithm. While the first method computes the extrinsic information based on the true probability density function (PDF) of the received signal, the second method relies on a nonlinear Gaussian approximation of the original PDF for its computation. Finally, the third relies on a description of the received signal by a linear model with a Gaussian additive distortion term.

Numerical results show that employing the common LLR computation method for AWGN channels without taking into account the effects of the discrete precoder causes an error floor in the systems' BER for high-SNR. By relying on more sophisticated LLR computation methods, the proposed approaches mitigate this problem while also enhancing the overall BER performance of the system. The contributions of the detection part were published in [36, 37].

1.4 Thesis Outline

The remainder of this thesis consists of 7 chapters which are structured as follows:

– Chapter 2 presents the model of the system;

- Chapter 3 revises the state-of-the-art and carefully describes the most important algorithms from literature;
- Chapter 4 presents the developed precoding techniques;
- Chapter 5 exposes the derivation of the proposed soft detection approaches as well as the DPA-IDD scheme;
- Chapter 6 presents the numerical evaluation of the proposed approaches and compares them with the techniques presented in chapter 3;
- Chapter 7 presents the conclusion of the thesis;
- Chapter 8 discusses possible extensions for the proposed study;

Finally, the convexity analysis and the derivation of the linear model are presented in the appendix.

1.5 Notation

Regarding the notation, bold lower case and upper case letters indicate vectors and matrices, non-bold letters express scalars. The operators $(\cdot)^*$, $(\cdot)^T$ and $(\cdot)^H$ stand for complex conjugation, transposition and Hermitian transposition, respectively. Real and imaginary part operator are also applied to vectors and matrices, e.g., $\mathrm{Re}\{\boldsymbol{x}\} = [\mathrm{Re}\{[\boldsymbol{x}]_1\}, \ldots, \mathrm{Re}\{[\boldsymbol{x}]_M\}]^T$ and equivalently for $\mathrm{Im}\{\cdot\}$.

2
System Model

This study considers a single-cell multiuser MIMO DL system where the BS has perfect channel state information (CSI) and is equipped with M transmitting antennas which serves K single-antenna users as illustrated by Fig. 2.1.

A blockwise transmission is considered in which the BS delivers N_b bits for each user. The user specific block is denoted by the vector $\boldsymbol{m}_k = [m_{k,1}, \ldots, m_{k,N_b}]$, where the index k indicates the k-th user. Each vector $\boldsymbol{m}_k$ is encoded into a codeword vector denoted by $\boldsymbol{c}_k = [c_{k,1}, \ldots, c_{k,\frac{N_b}{R}}]$, where R is the code rate. The encoding operation is considered to be systematic meaning that $\boldsymbol{c}_k = [p_{k,1}, \ldots, p_{k,\frac{N_b(1-R)}{R}}, m_{k,1}, \ldots m_{k,N_b}]$, where $p_{k,i}$ is the i-th parity bit.

Each encoder provides, sequentially over time slots, N bits to a modulator which maps them into a symbol $s \in \mathcal{S}$ using Gray coding. The set $\mathcal{S}$ represents all possible symbols of a α_s-PSK modulation, with $\alpha_s = 2^N$, and is described by

$$\mathcal{S} = \left\{ s : s = e^{\frac{j\pi(2i+1)}{\alpha_s}}, \text{ for } i = 1, \ldots, \alpha_s \right\}. \tag{2-1}$$

The mapping operation is denoted as $s_k[t] = \mathcal{M}(\boldsymbol{r}_{k,t})$, where $\boldsymbol{r}_{k,t} = [r_{k,t,1}, \ldots, r_{k,t,N}]$ is the t-th bit vector, taken from $\boldsymbol{c}_k$. The vector $\boldsymbol{r}_{k,t}$ can also be expressed as $\boldsymbol{r}_{k,t} = [c_{k,(t-1)N+1}, \ldots, c_{k,tN}]$ for $t = 1, \ldots, \frac{N_b}{RN}$. After mapping, the symbols of all K users are represented in a stacked vector notation as $\boldsymbol{s}[t] = [s_1[t] \ldots s_K[t]]^T \in \mathcal{S}^K$ for each time slot t.

The vector $\boldsymbol{s}[t]$ is forwarded to the precoder, which computes the transmit vector $\boldsymbol{x}[t] = [x_1[t] \ldots x_M[t]]^T \in \mathcal{X}^M$. The entries of the transmit vector are constrained to the set $\mathcal{X}$, which describes an α_x-PSK alphabet, denoted by

$$\mathcal{X} = \left\{ x : x = e^{\frac{j\pi(2i+1)}{\alpha_x}}, \text{ for } i = 1, \ldots, \alpha_x \right\}. \tag{2-2}$$

A frequency flat fading channel is considered, which is described by the matrix $\boldsymbol{H}$ with coefficients $h_{k,m} = g_{k,m}\sqrt{\beta_k}$ where $g_{k,m}$ represents the complex small-scale fading between the m-th antenna and the k-th user, β_k denotes the real valued large-scale fading coefficient of the k-th user, $k = 1...K$ and $m = 1...M$. A block fading model is considered in which the channel is time-

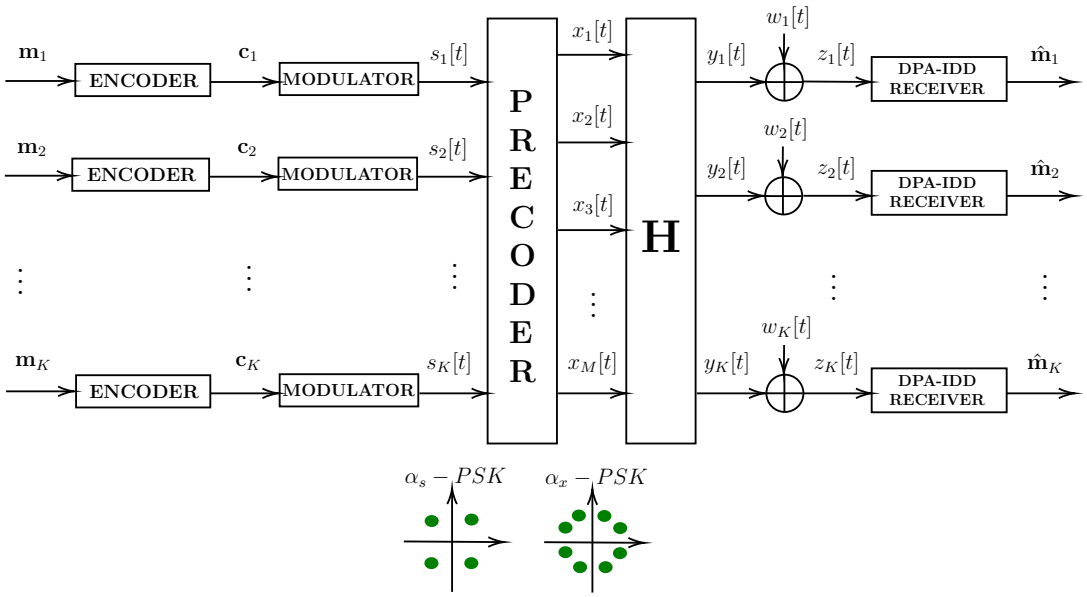

Figure 2.1: Multiuser MIMO downlink with discrete precoding and channel coding

invariant during the transmission time.

The BS computes for each coherence time interval of the channel the lookup-table $\mathcal{L}$ containing all possible precoding vectors, which then implies $\boldsymbol{s} \in \mathcal{S}^K \iff \boldsymbol{x}(\boldsymbol{s}) \in \mathcal{L}$.

With this, the noiseless received signal corresponding to the k-th user at the t-th time slot is given by

$$y_k[t] = \boldsymbol{h}_k \, \boldsymbol{x}[t], \tag{2-3}$$

where $\boldsymbol{h}_k$ is the k-th row of the channel matrix $\boldsymbol{H}$. At the user terminals the noiseless received signals are distorted by AWGN denoted by the complex random variable $w_k[t] \sim \mathcal{CN}(0, \sigma_w^2)$. The receive signal from the k-th user is denoted by

$$\begin{aligned} z_k[t] &= y_k[t] + w_k[t] \\ &= \boldsymbol{h}_k \, \boldsymbol{x}[t] + w_k[t], \end{aligned} \tag{2-4}$$

Using stacked vector notation equation (2-4) can be extended to

$$\begin{aligned} \boldsymbol{z}[t] &= \boldsymbol{y}[t] + \boldsymbol{w}[t] \\ &= \boldsymbol{H} \, \boldsymbol{x}[t] + \boldsymbol{w}[t], \end{aligned} \tag{2-5}$$

where $\boldsymbol{z}[t] = [z_1[t] \dots z_K[t]]^T$, $\boldsymbol{y}[t] = [y_1[t] \dots y_K[t]]^T$ and $\boldsymbol{w}[t] = [w_1[t] \dots w_K[t]]^T$.

Each received signal $z_k[t]$ is forwarded to the IDD receiver where the transmitted block will be detected. Finally the data block available to the k-th user reads as $\hat{\boldsymbol{m}}_k = [\hat{m}_{k,1}, \ldots, \hat{m}_{k,N_b}]$.

3
Literature Review and Important Baselines

In this chapter, the precoding literature and important precoding baselines for MIMO DL systems is reviewed. The precoding techniques present in the literature can be divided into three major groups, they are

- Total Power Constraint (TPC),
- Per Antenna Power Constraint (PAPC),
- Low Resolution.

This thesis will focus primarily on the development of low resolution precoders. However, in what follows a brief explanation on TPC and PAPC precoders will be also given.

3.1
Total Power Constraint Precoding

This subsection describes the TPC precoding technique. This type of precoder is characterized by the TPC, meaning that the norm squared of the precoding vector must be equal to the total transmit power available. The objective is then to minimize the BER subject to the TPC. There are several criteria for BER minimization available for precoding, but, in general, the TPC optimization problem can be cast as

$$\begin{aligned} \min_{\boldsymbol{x}[t]} \quad & f(\boldsymbol{H}, \boldsymbol{x}[t]) \\ \text{s.t.} \quad & \mathrm{E}\left[||\boldsymbol{x}[t]||_2^2\right] \leq \mathrm{E}_{\mathrm{tx}} \ , \end{aligned} \tag{3-1}$$

where $f(\boldsymbol{H}, \boldsymbol{x}[t])$ is the objective function that represents the design criterion and E_{tx} is the maximum transmit power.

The TPC strategy generates the higher BER performance of all classes of precoding techniques. However, according to [18] the assumption of a TPC is not realistic, since each transmitting antenna is typically characterized by its own PA and is hence affected by specific power constraints. Moreover, the usage of precoding techniques where the power at each antenna is fixed allows the PA to work with minimum peak-to-average ratio, hence enhancing the EE

of the system. Next, the popular linear MMSE Precoder is introduced as an example of TPC precoding technique.

3.1.1 Linear MMSE Precoder

As the name states the objective of the linear MMSE precoder is to minimize the mean square error between the arriving signal $\boldsymbol{z}[t]$ and the data $\boldsymbol{s}[t]$ subject to the TPC. The noise and signal statistics are supposed to be known as well as the channel state. The problem formulation is then,

$$\begin{aligned} \boldsymbol{P}_{\text{MMSE}} = \arg\min_{\boldsymbol{P},f} \ & \text{E}\left[||\boldsymbol{s}[t] - f\boldsymbol{z}[t]||_2^2\right] \\ \text{s.t.} \quad & \text{tr}\left\{\boldsymbol{P}\boldsymbol{C}_s\boldsymbol{P}^H\right\} \leq \text{E}_{\text{tx}} \ . \end{aligned} \tag{3-2}$$

Note that the constraint $\text{tr}\left\{\boldsymbol{P}\boldsymbol{C}_s\boldsymbol{P}^H\right\} \leq \text{E}_{\text{tx}}$ is equivalent to $\text{E}\left[||\boldsymbol{x}||_2^2\right] = \text{E}_{\text{tx}}$. Considering the MSE minimization problem in (3-2) one might get the idea that an equivalent problem can be cast when scaling the symbol vector $\boldsymbol{s}$ instead of the received signal $\boldsymbol{z}$. This case however implies that issue that scaling the vector $\boldsymbol{s}$ yields a scaled MSE where the scaling factor is part of the optimization problem.

Solving the optimization problem from (3-2) yields

$$\boldsymbol{P}_{\text{MMSE}} = f_{\text{MMSE}}\left(\boldsymbol{H}^H\boldsymbol{H} + \frac{\text{tr}\left\{\boldsymbol{C}_w\right\}}{\text{E}_{\text{tx}}}\boldsymbol{I}\right)^{-1}\boldsymbol{H}^H \tag{3-3}$$

$$f_{\text{MMSE}} = \sqrt{\frac{\text{E}_{\text{tx}}}{tr\left\{\left(\boldsymbol{H}^H\boldsymbol{H} + \frac{\text{tr}\{\boldsymbol{C}_w\}}{\text{E}_{\text{tx}}}\boldsymbol{I}\right)^{-2}\boldsymbol{H}^H\boldsymbol{C}_s\boldsymbol{H}\right\}}} \ , \tag{3-4}$$

where $\boldsymbol{C}_w$ is the covariance matrix of the noise and f_{MMSE} is the scaling factor that guarantees that the total transmit power is constant and less or equal to the total power constraint E_{tx}.

3.2 Per Antenna Power Constraint Precoding

This subsection describes PAPC precoders. This group of precoders is characterized by the PAPC meaning that the absolute value squared of each entry of the precoder's transmit vector has the same value. Essentially this results in equal distribution of power between all transmit antennas. In general

the PAPC optimization problem can be cast as

$$\begin{aligned} \min_{\boldsymbol{x}[t]} \quad & f(\boldsymbol{H}, \boldsymbol{x}[t]) \\ \text{s.t.} \quad & |x_i[t]|^2 = \frac{\mathrm{E_{tx}}}{M} \quad \text{for} \quad i = 1 \ ... \ M. \end{aligned} \tag{3-5}$$

Precoding using PAPC leads to more efficient usage of PAs since it allows the PA to work with the minimum the peak-to-average ratio. However, PAPC precoding does not address the problem of energy efficient DACs since it still implies high resolution.

3.3 Low Resolution Precoding

This subsection describes low resolution precoders. In this type of system a new constraint takes place of the PAPC, now each entry of the transmit vector belongs to a predefined discrete set. This set generally represents a PSK modulation, however, some precoders have extended the PSK results for QAM. However, when QAM modulation is used, the PA's efficiency goes down, decreasing the system's EE. Because of the discrete set constraint, the term discrete precoding is treated as a synonym of low-resolution precoding.

Having a discrete set of precoding symbol options generates a significantly harder problem in terms of computational complexity when compared with the classes presented in subsections 3.1 and 3.2. In general the problem of low resolution precoding states

$$\begin{aligned} \min_{\boldsymbol{x}[t]} \quad & f(\boldsymbol{H}, \boldsymbol{x}[t]) \\ \text{s.t.} \quad & \boldsymbol{x}[t] \in \mathcal{F} \\ & \mathrm{E}\left[||\boldsymbol{x}[t]||_2^2\right] \leq \mathrm{E_{tx}}, \end{aligned} \tag{3-6}$$

where $\mathcal{F}$ is the discrete set. Note that in general the low-resolution precoding optimization problem implies the TPC. However, depending on the structure of the set $\mathcal{F}$ the TPC becomes redundant and, thus, can be dropped.

3.4 Low Resolution Precoding With PSK Modulation

One interesting result happens when the discrete set represents a PSK modulation. In this case, not only the TPC can be dropped, but also the set $\mathcal{F}$ is in the form of $\mathcal{X}^M$. As will be seen in Section 4.1 this is useful for tree search

based methods. When the entries of transmit vector $\boldsymbol{x}[t]$ are constrained to a PSK modulation the optimization problem in (3-6) can be rewritten as

$$\begin{aligned} \min_{\boldsymbol{x}[t]} \quad & f(\boldsymbol{H}, \boldsymbol{x}[t]) \\ \text{s.t.} \quad & \boldsymbol{x}[t] \in \mathcal{X}^M. \end{aligned} \tag{3-7}$$

Most of state-of-the-art algorithms, for solving (3-7), employ relaxations of the feasible set to convexify the non convex optimization problem.

When a relaxation is performed, the solution of the relaxed version of (3-7) is a lower bound on the optimal solution. However, this lower bound solution does not necessarily belong in $\mathcal{X}^M$ therefore being considered unfeasible on the original problem. For finding a feasible solution from (3-7) in general mapping to the closest Euclidean distance point in $\mathcal{X}^M$ is considered.

Yet, this mapping step does not preserve the optimality, since the point in $\mathcal{X}^M$ that is closest to the lower bound solution is not necessarily the optimal of (3-7).

Most of the state-of-the-art low resolution precoders employ the mentioned mapping step and, thus, are not optimal in their design criterion. In the following, some of the most popular state-of-the-art low resolution precoders are presented.

3.4.1 Phase Zero-Forcing Mapped Precoder

The Phase Zero-Forcing mapped precoder (ZF-P) is a low complexity approach presented by [11]. It considers the popular TPC ZF precoder presented by [13] and utilizes it in the low resolution context.

The idea is simply to multiply the data vector $\boldsymbol{s}[t]$ by the ZF precoder matrix and scale the result such that the PAPC is met. The mathematical expression is stated as follows

$$\boldsymbol{x}[t] = \sqrt{\frac{\mathrm{E_{tx}}}{M}} \; e^{j \arg\{\boldsymbol{P}_{\mathrm{zf}} \boldsymbol{s}[t]\}} \;, \tag{3-8}$$

where

$$\begin{aligned} \boldsymbol{P}_{\mathrm{zf}} &= f_{\mathrm{zf}} \boldsymbol{H} \left(\boldsymbol{H}^H \boldsymbol{H} \right)^{-1} \\ f_{\mathrm{zf}} &= \sqrt{\frac{\mathrm{E_{tx}}}{\mathrm{tr}\left\{ \boldsymbol{C}_s \left(\boldsymbol{H}\boldsymbol{H}^H \right)^{-1} \right\}}} \;, \end{aligned} \tag{3-9}$$

is the TPC ZF precoding matrix.

Note that the result from equation (3-8) is not necessarily in $\mathcal{X}^M$ and, thus, the discussed mapping step is considered to generate a feasible solution.

3.4.2 CVX-CIO Mapped Precoder

The CVX-CIO precoder, presented in [18], is an algorithm based on the maximization of the minimum value of constructive interference. For finding the transmit vector a second order cone program (SOCP) is solved subject to the PAPC.

The general idea of CVX-CIO is that the signal that arrives at the receiver is the transmitted signal summed with noise and interference. Since the BS is assumed to have perfect CSI, the interference can be controlled in such a way that it can be helpful for detection.

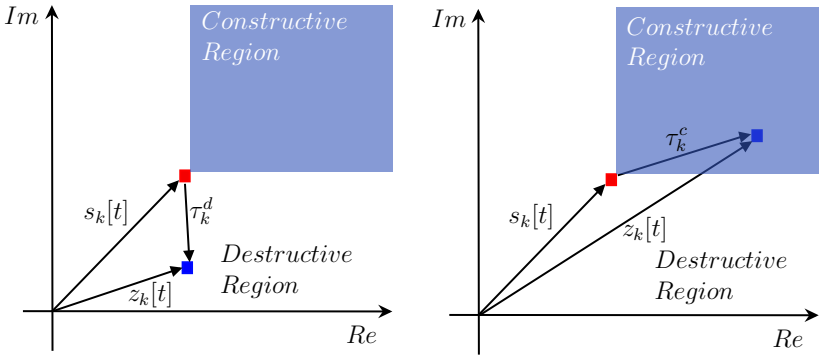

Figure 3.1: Destructive (left) and Constructive (right) interference in for QPSK data system

A graphical representation of the concept can be seen in Fig. 3.1. The constructive interference of the k-th user is denoted τ_k^c and the destructive interference is denoted τ_k^d. Since the system depicted in Fig. 3.1 utilizes QPSK modulation, the decision thresholds are the real and imaginary axis. Note that when the interference is constructive the received symbol $z_k[t]$ is more distant to the decision thresholds than $s_k[t]$. However, when the interference is destructive the received symbol $z_k[t]$ is closer to the decision thresholds than $s_k[t]$.

Using the concept of constructive and destructive interference the work from [18] defines a metric for optimizing the minimum value of constructive interference. The optimization problem derived with this metric is stated as

follows

$$\begin{aligned} \boldsymbol{x}[t] &= \arg\max \left\{ \mathrm{Re}\left\{\tau_k\right\} \tan\psi - \left|\mathrm{Im}\left\{\tau_k\right\}\right| \right\} \\ &\text{s.t.} \quad |x_i[t]|^2 = \frac{\mathrm{E_{tx}}}{M} \quad \text{for} \quad i = 1 \ ... \ M. \end{aligned} \tag{3-10}$$

where,

$$\psi = \frac{\pi}{\alpha_s} \ , \quad \tau_k = \left(\sum_{m=1}^{M} h_{k,m} x_m[t] - s_k[t] \right) e^{-j\phi_k} \ , \quad \phi_k = \arg\left\{ s_k[t] \right\} \ . \tag{3-11}$$

Note that, since the feasible set described by the PAPC is not convex, the optimization problem from (3-10) is also not convex. To convexify the mentioned problem the PAPC is relaxed to an inequality. The relaxed optimization problem reads as

$$\boldsymbol{x}'[t] = \arg \max_{|\boldsymbol{x}'[t]| \leq \sqrt{M}} \left\{ \mathrm{Re}\left\{\tau_k\right\} \tan\psi - \left|\mathrm{Im}\left\{\tau_k\right\}\right| \right\}. \tag{3-12}$$

The optimal precoding vector computed from (3-12) does not necessarily yield feasible solution of the optimization problem from (3-10). Therefore, for finding the transmit vector, the solution from (3-12) needs to be scaled elementwise as seen below

$$x_i[t] = \sqrt{\frac{\mathrm{E_{tx}}}{M}} \frac{x'_i[t]}{|x'_i[t]|} \ \text{for} \ \ i = 1 \ ... \ M. \tag{3-13}$$

After scaling, mapping is considered to find a solution in the feasible set $\mathcal{X}^M$.

3.4.3 RedMinBER Precoder

The Reduced Dimension Minimum BER (RedMinBER) precoder [22] is a suboptimal massive MIMO approach based on the MMDDT criterion (the MMDDT criterion is detailed in section 4.2) for 1-bit quantized transmit signals.

As previously mentioned, a common method for developing suboptimal low-resolution precoders is to describe the problem in the form of (3-7), employ a relaxation of the feasible set and subsequently map the solution to $\mathcal{X}^M$.

However, the RedMinBER precoder employs a different technique for computing $\boldsymbol{x}[t]$. The idea is to assume that $\boldsymbol{x}[t] = \mathcal{Q}\left(\boldsymbol{P}\left(\boldsymbol{s}[t] + \boldsymbol{\epsilon}\right)\right)$, where $\boldsymbol{P}$ is a fixed linear precoder that only depends on the channel state (e.g. the ZF precoder) , $\boldsymbol{\epsilon}$ is a predistortion term, and $\mathcal{Q}(\cdot)$ is the 1-bit quantizer operation.

The advantage is that the optimization will be made in $\boldsymbol{\epsilon}$ which has dimension $K \times 1$. Since [22] considers a system where $K \ll M$ the computational complexity of solving the optimization problem is reduced.

The approach from [22] then utilizes a gradient descent method for updating $\boldsymbol{\epsilon}$ and maximizing the minimum distance to the decision threshold, which is given by

$$\delta = \arg \max_{\boldsymbol{x} \in \mathcal{X}^M} \min_k \delta_k, \tag{3-14}$$

where

$$\delta_k = \mathrm{Re}\left\{z_k[t]\right\} \sin\theta - \left|\mathrm{Im}\left\{z_k[t]\right\}\right| \cos\theta, \quad \theta = \frac{\pi}{\alpha_s}, \tag{3-15}$$

and $k = 1, \ldots, K$. At the p-th iteration the gradient, denoted by $\tilde{\nabla}_{\boldsymbol{\epsilon}}$, is given by

$$\tilde{\nabla}_{\boldsymbol{\epsilon}} \delta \approx \boldsymbol{P}^H \tilde{\boldsymbol{H}} \boldsymbol{e}_k \left(\sin\theta + j \ \mathrm{sign}\left(z_{kI} \cos\theta\right)\right), \tag{3-16}$$

where

$$\tilde{\boldsymbol{H}} = \left(\mathrm{diag}\left(\boldsymbol{s}[t]\right)\right)^H \boldsymbol{H}, \quad z_{kI} = \mathrm{Im}\left\{z_k[t]\right\}, \tag{3-17}$$

$\boldsymbol{e}_k$ is a vector with a one in position k and zeros elsewhere and the index k denotes the user with the smallest distance to the decision threshold δ_k. With this, the steps of the algorithm of the RedMinBER precoder are detailed in Algorithm 1. Note that, the transmit vector $\boldsymbol{x}[t]$ is computed using the output solution of Algorithm 1 as $\boldsymbol{x}[t] = \mathcal{Q}\left(\boldsymbol{P}\ \boldsymbol{s}_{\mathrm{opt}}\right)$.

Algorithm 1 RedMinBER Precoder's Algorithm

Given $\boldsymbol{s}[t]$, $\tilde{\boldsymbol{H}}$, $\boldsymbol{P}$, number of iterations N_p and stepsize μ, set $p = 1$ and $\epsilon(1) = 0$
Calculate $\boldsymbol{z} = \tilde{\boldsymbol{H}} \mathcal{Q}\left(\boldsymbol{P}\boldsymbol{s}[t]\right)$ and $\delta(1)$ using equations (3-14) and (3-15)
Set $\boldsymbol{s}_{\mathrm{opt}} = \boldsymbol{s}[t]$ and $\delta_{\mathrm{opt}} = \delta(1)$
for $p = 1 : N_p$ **do**
  Find $\boldsymbol{\epsilon}(p+1) = \epsilon(p) + \mu \tilde{\nabla}^*_{\boldsymbol{\epsilon}} \delta(p)$
  Calculate $\boldsymbol{z} = \tilde{\boldsymbol{H}} \mathcal{Q}\left(\boldsymbol{P}\left(\boldsymbol{s}[t] + \boldsymbol{\epsilon}(p+1)\right)\right)$ and $\delta(p+1)$
  If $\delta(p+1) > \delta_{\mathrm{opt}}$ set $\delta_{\mathrm{opt}} = \delta(p+1)$ and $\boldsymbol{s}_{\mathrm{opt}} = \boldsymbol{s}[t] + \boldsymbol{\epsilon}(p+1)$
end for
Output solution $\boldsymbol{s}_{\mathrm{opt}}$

The main advantage of using the RedMinBER algorithm is its reduced computational complexity for systems where $K \ll M$. As shown in the study of [22], the BER performance of the RedMinBER method is similar to the

performance of the CVX-CIO approach presented in the previous subsection when the ZF method is employed as the linear precoding method.

3.4.4 MSM Precoder

The maximum safety margin (MSM) precoder [20] is a suboptimal algorithm based on the MMDDT criterion. The MSM algorithm is suitable for PSK modulations in which $\alpha_x = 2^q$ where $q \in \mathbb{N}$ is the number of possible outputs of the phase quantizer.

The MSM precoder relies on the relation of the original discrete set $\mathcal{X}^M$ to its convex hull. By doing that the non convex problem turns into convex permitting solution via standard convex optimization algorithms. The optimization problem in which the MSM algorithm is based follows

$$\begin{aligned} &\max_{\boldsymbol{x}[t],\delta} \begin{bmatrix} \mathbf{0}_{2M}^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}[t] \\ \delta \end{bmatrix} \quad \text{s.t.} \\ &\begin{bmatrix} \boldsymbol{B} - \tan(\theta)\boldsymbol{A} & \frac{1}{\cos\theta}\mathbf{1}_K \\ -\boldsymbol{B} - \tan(\theta)\boldsymbol{A} & \frac{1}{\cos\theta}\mathbf{1}_K \\ \boldsymbol{E} & \mathbf{0}_{M(\alpha_s-4)} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}[t] \\ \delta \end{bmatrix} \leq \begin{bmatrix} \mathbf{0}_{2K} \\ \cos\left(\frac{\pi}{\alpha_s}\mathbf{1}_{2M}\right) \end{bmatrix} \\ &\text{and} \quad \begin{bmatrix} -\cos\left(\frac{\pi}{\alpha_s}\mathbf{1}_{2M}\right) \\ 0 \end{bmatrix} \leq \begin{bmatrix} \boldsymbol{x}[t] \\ \delta \end{bmatrix} \leq \begin{bmatrix} \cos\left(\frac{\pi}{\alpha_s}\mathbf{1}_{2M}\right) \\ 0 \end{bmatrix}, \end{aligned} \tag{3-18}$$

where

$$\begin{aligned} &\boldsymbol{H}_{s^*} = \mathrm{diag}(\boldsymbol{s}^*[t])\boldsymbol{H}, \quad \boldsymbol{A} = \begin{bmatrix} \mathrm{Re}\{\boldsymbol{H}_{s^*}\} & \mathrm{Im}\{\boldsymbol{H}_{s^*}\} \end{bmatrix}, \\ &\boldsymbol{B} = \begin{bmatrix} \mathrm{Im}\{\boldsymbol{H}_{s^*}\} & \mathrm{Re}\{\boldsymbol{H}_{s^*}\} \end{bmatrix}, \quad \boldsymbol{R}_i = \begin{bmatrix} \cos\beta_i & \sin\beta_i \\ -\sin\beta_i & \cos\beta_i \end{bmatrix}, \\ &\beta_i = \frac{2\pi}{\alpha_x}(i-1), \quad \boldsymbol{E} = \begin{bmatrix} \boldsymbol{R}_2^T & -\boldsymbol{R}_2^T & \cdots \boldsymbol{R}_{\frac{\alpha_s}{4}}^T & -\boldsymbol{R}_{\frac{\alpha_s}{4}}^T \end{bmatrix}, \quad \theta = \frac{\pi}{\alpha_s}. \end{aligned}$$

Solving (3-18) does not necessarily provide a feasible solution for the original problem. Therefore, after solving, the resulting vector $\boldsymbol{x}[t]$ must be mapped into the feasible set $\mathcal{X}^M$ in order to achieve a feasible solution of the problem in (3-18).

Finally, it is worth noting that for the special case where the transmit vector is one-bit quantized ($\alpha_x = 4$) another precoder based on the same formulation of the objective function was developed. The MSM-FISTA precoder [23] is a low complexity one-bit approach that provides lower BER when compared with the MSM precoder especially for systems with denser data modulation.

3.4.5 Precoding Using Exhaustive Search

The Exhaustive Search method is an optimal high complexity algorithm that provides the transmit vector by solving the non convex problem (3-7) by evaluating all possibilities.

This algorithm is very flexible in the sense that any criterion can be used. However, the most popular and commonly employed for this method are the MMDDT and the MSE. In that sense if, for example, MSE is chosen, the Exhaustive Search precoder will exploit all possible points in the set $\mathcal{X}^M$ and find $\boldsymbol{x}[t]$ that minimizes the mean square error from the arriving symbol and the sent data.

This method is optimal in the sense that it always provides the optimal vector given a chosen criterion, which also usually provides a lower uncoded BER. Yet, the complexity of Exhaustive Search is given by $\mathcal{O}\left(\alpha_x^M\right)$ which is prohibitive for high values of modulation order or number of transmit antennas.

4
Optimal Discrete Precoding via Branch-and-Bound

This chapter exposes the derivation of the developed precoding techniques. Throughout it, the objective functions for the precoding problem will be explained as well as its design criteria and problem formulation.

The main goal of MIMO precoding algorithms is to mitigate the multiuser interference (MUI) and to simultaneously reduce distortions brought by additive noise. As already mentioned in the case of precoding with low-resolution the problem is more difficult than with full resolution because the transmitting vector is, in this case, constrained to a discrete set.

The rest of the chapter is divided into four sections. The first provides a brief introduction to the branch-and-bound method, while the second and third expose the derivation of the proposed MMDDT and MMSE precoding strategies, respectively. The fourth and final section provides the final considerations on the subject. The proposed precoding methods do not rely on previous time instants to compute the transmit vector. For this reason, the index t notation is suppressed in this section.

4.1
Introduction of the Branch-and-Bound Method

The branch-and-bound method, first created in 1960 by A. H. Land and A. G. Doig [38], is an established technique in the wireless communications area. It has important applications in this context, such as multiuser detection [39], discrete beamforming [40, 41] and more recently discrete precoding [25].

A branch-and-bound algorithm is a tree search based method. The tree represents the set of all possible solutions for the vector $\boldsymbol{x}$, i.e., it is a representation of the set $\mathcal{X}^M$. For the construction of the tree M levels are considered and each node has one ingoing branch and α_x outgoing branches as shown in Fig. 4.1.

For constructing the precoding vector the minimization of an objective function $g(\boldsymbol{x}, \boldsymbol{s})$ subject to the feasible discrete set is considered. The corresponding optimization problem is described as

$$\boldsymbol{x}_{\text{opt}} = \arg\min_{\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{s}) \quad \text{s.t. } \boldsymbol{x} \in \mathcal{X}^M. \tag{4-1}$$

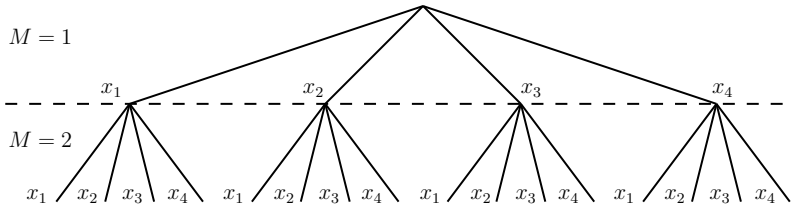

Figure 4.1: Tree representation of the set $\mathcal{X}^M$ for a system with $M = 2$ BS antennas and QPSK precoding modulation ($\alpha_x = 4$)

A lower bound on $g(\boldsymbol{x}_{\text{opt}}, \boldsymbol{s})$ can be obtained by relaxing $\mathcal{X}^M$ to its convex hull. The relaxed problem is expressed as

$$\boldsymbol{x}_{\text{lb}} = \arg\min_{\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{s}) \quad \text{s.t. } \boldsymbol{x} \in \mathcal{P}. \tag{4-2}$$

An associated upper bound on $g(\boldsymbol{x}_{\text{opt}}, \boldsymbol{s})$ can be obtained by mapping the solution of (4-2) and evaluating $g(\cdot)$ accordingly. The upper bound value of (4-1) is termed $\check{g}$.

Having an upper bound solution implies that $\check{g} \geq g(\boldsymbol{x}_{\text{opt}}) \geq g(\boldsymbol{x}_{\text{lb}})$, which means that the mapped solution is always greater or equal to the relaxed one from (4-2).

By fixing d entries of $\boldsymbol{x}$, the vector can be rewritten as $\boldsymbol{x} = [\boldsymbol{x}_1^T, \boldsymbol{x}_2^T]^T$, with $\boldsymbol{x}_1 \in \mathcal{X}^d$. With this, a subproblem can be formulated as

$$\begin{aligned} \boldsymbol{x}_2 &= \arg\min_{\boldsymbol{x}_2} g(\boldsymbol{x}_2, \boldsymbol{x}_1, \boldsymbol{s}) \\ &\text{s.t. } \boldsymbol{x}_2 \in \mathcal{X}^{M-d}. \end{aligned} \tag{4-3}$$

Relaxing the problem from (4-3) states

$$\begin{aligned} \boldsymbol{x}_{2,\text{lb}} &= \arg\min_{\boldsymbol{x}_2} g(\boldsymbol{x}_2, \boldsymbol{x}_1, \boldsymbol{s}) \\ &\text{s.t. } \boldsymbol{x}_2 \in \mathcal{J}, \end{aligned} \tag{4-4}$$

where $\mathcal{J}$ is the convex hull of $\mathcal{X}^{M-d}$.

If the optimal value of (4-4) is larger than a known upper bound $\check{g}$ on the solution of (4-1), then all members in the discrete set which include the previously fixed vector $\boldsymbol{x}_1$ can be excluded from the search process.

The purpose of this strategy is to exclude most of the candidates from the possible solution set, such that the number of residual candidates is only a small fraction of its total number and, thus, they can be evaluated via exhaustive search.

4.2 Proposed Optimal MMDDT Precoder Design

This section establishes the objective of the MMDDT precoding algorithm and exposes the problem formulation.

The aim of the precoding strategy is to find the vector $\boldsymbol{x}$ which yields the noiseless received vector $\boldsymbol{y}$ where the smallest distance to the decision threshold is maximized. The objective can be expressed in the epigraph form [42], which then corresponds to a linear objective function with linear constraints. Taking into account the quantization at each transmit antenna, the feasible set is discrete, which then yields a non-convex problem.

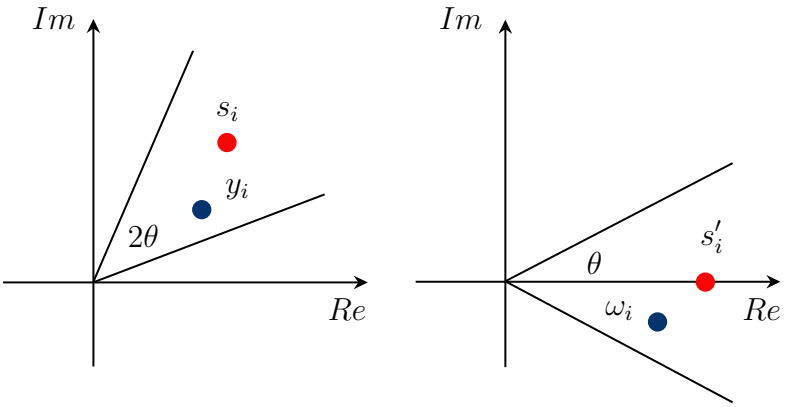

Figure 4.2: Rotated coordinate system

During this section the minimum distance to the decision threshold is denoted by ϵ. By considering a rotation by $\arg\{s_i^*\} = -\phi_{s_i}$ of the coordinate system, the symbol of interest is placed on the real axis, as shown in Fig.4.2. This is done by multiplying both the interest symbol s_i and the noiseless received signal y_i by $e^{-j\phi_{s_i}} = s_i^*$ which reads

$$s'_i = s_i s_i^* = 1, \quad \omega_i = y_i s_i^*. \tag{4-5}$$

The smallest distance of the rotated symbol ω_i to the rotated decision threshold is then expressed as

$$\epsilon_i = \mathrm{Re}\{\omega_i\} \sin\theta - |\mathrm{Im}\{\omega_i\}| \cos\theta, \tag{4-6}$$

where $\theta = \frac{\pi}{\alpha_s}$. This description is shown in more details in [20]. Since the considered rotation includes also the decision thresholds, the distance expression in (4-6) holds also for y_i. The minimum of all ϵ_i, for $i = 1, \ldots, K$ is defined as ϵ, which serves as the objective of the precoding design. The

objective of the algorithm is to construct the transmit vector $\boldsymbol{x}$ that maximizes ϵ.

Based on a stacked vector notation for ω_i, namely $\boldsymbol{\omega} = \mathrm{diag}(\boldsymbol{s}^*)\boldsymbol{H}\boldsymbol{x}$, the equivalent minimization problem reads as

$$\begin{aligned}\left[\boldsymbol{x}_{\mathrm{opt}},\, \epsilon_{\,\mathrm{opt}}\right] = \underset{\boldsymbol{x}\in\mathcal{X}^M,\epsilon}{\mathrm{argmin}}\ & -\epsilon \\ \text{s.t.}\ \ & \mathrm{Re}\left\{\boldsymbol{H}_{s^*}\boldsymbol{x}\right\}\sin\theta - \left|\mathrm{Im}\left\{\boldsymbol{H}_{s^*}\boldsymbol{x}\right\}\right|\cos\theta \geq \epsilon\mathbf{1}_{2K},\end{aligned} \tag{4-7}$$

where $\boldsymbol{H}_{s^*} = \mathrm{diag}(\boldsymbol{s}^*)\boldsymbol{H}$. Note that, the description of the objective is equivalent to the one presented in [20] and for the special case of QPSK modulation is also analogous to the one utilized in [25].

4.2.1 MMDDT-Mapped Precoder

One approach for finding a feasible solution of (4-7) is to solve a relaxed version of the original problem followed by a mapping process to ensure that the precoding vector is in the feasible set of the discrete problem.

The relaxation is brought by replacing the set $\mathcal{X}^M$ by its convex hull, which then establishes convexity of the considered problem. The corresponding relaxed problem reads as

$$\begin{aligned}\left[\boldsymbol{x}_{\mathrm{lb}},\, \epsilon_{\mathrm{lb}}\right] = \arg\min_{\boldsymbol{x},\epsilon}\ & -\epsilon \\ \text{s.t.}\ \ & \mathrm{Re}\left\{\boldsymbol{H}_{s^*}\boldsymbol{x}\right\}\sin\theta - \left|\mathrm{Im}\left\{\boldsymbol{H}_{s^*}\boldsymbol{x}\right\}\right|\cos\theta \geq \epsilon\mathbf{1}_{2K} \\ & \mathrm{Re}\left\{x_m e^{j\phi_i}\right\} \leq \frac{\cos\left(\frac{\pi}{\alpha_x}\right)}{\sqrt{M}},\ \text{for } m = 1,\ldots,M \text{ and} \\ & \phi_i = \frac{2\pi i}{\alpha_x},\ \text{for } i = 1,\ldots\alpha_x,\end{aligned} \tag{4-8}$$

which is basically presented before in [20]. With the equivalent real valued notation, the problem in (4-8) can be expressed as a linear program (LP). Note that unlike the algorithm in [20], where α_x is restricted to integer powers of 2, the problem formulation (4-8) from above supports α_x to be any integer value. Subsequently the continuous solution $\boldsymbol{x}_{\mathrm{lb}}$ is quantized to the point in $\mathcal{X}^M$ with the shortest Euclidean distance.

The optimal value of (4-8) is always a lower bound to the optimal value of the original problem (4-7). Mapping to the feasible set yields a valid solution $\boldsymbol{x}_{\mathrm{ub}}$ and the corresponding value for $-\epsilon$ provides an upper bound on the optimal value of the original problem (4-7).

4.2.2 MMDDT Branch-and-Bound Algorithm Derivation

In this section a branch-and-bound algorithm is proposed which solves (4-7) by considering the problem in (4-8) for the initialization and the sub-problems strategy given by (4-3) for computing lower bounds.

In order to formulate a real valued problem matrix $\boldsymbol{H}_r$ and vector $\boldsymbol{x}_r$ are defined as follows

$$\boldsymbol{x}_{\mathrm{r}} = \begin{bmatrix} \mathrm{Re}\{\boldsymbol{x}_1\} \\ \mathrm{Im}\{\boldsymbol{x}_1\} \\ \mathrm{Re}\{\boldsymbol{x}_2\} \\ \mathrm{Im}\{\boldsymbol{x}_2\} \\ \vdots \\ \mathrm{Re}\{\boldsymbol{x}_M\} \\ \mathrm{Im}\{\boldsymbol{x}_M\} \end{bmatrix}, \quad \boldsymbol{H}_{\mathrm{r}} = \begin{bmatrix} \Gamma_{11} \cdots \Gamma_{1M} \\ \Lambda_{11} \cdots \Lambda_{1M} \\ \vdots \\ \Gamma_{K1} \cdots \Gamma_{KM} \\ \Lambda_{K1} \cdots \Lambda_{KM} \\ \Psi_{11} \cdots \Psi_{1M} \\ \Delta_{11} \cdots \Delta_{1M} \\ \vdots \\ \Psi_{K1} \cdots \Psi_{KM} \\ \Delta_{K1} \cdots \Delta_{KM} \end{bmatrix}, \tag{4-9}$$

with

$$\begin{aligned} \boldsymbol{\Gamma} = &\quad \mathrm{Im}\{\boldsymbol{H}_{s^*}\}\cos(\theta) - \mathrm{Re}\{\boldsymbol{H}_{s^*}\}\sin(\theta) \\ \boldsymbol{\Lambda} = &\quad \mathrm{Re}\{\boldsymbol{H}_{s^*}\}\cos(\theta) + \mathrm{Im}\{\boldsymbol{H}_{s^*}\}\sin(\theta) \\ \boldsymbol{\Psi} = &-\mathrm{Im}\{\boldsymbol{H}_{s^*}\}\cos(\theta) - \mathrm{Re}\{\boldsymbol{H}_{s^*}\}\sin(\theta) \\ \boldsymbol{\Delta} = &\quad \mathrm{Im}\{\boldsymbol{H}_{s^*}\}\sin(\theta) - \mathrm{Re}\{\boldsymbol{H}_{s^*}\}\cos(\theta). \end{aligned} \tag{4-10}$$

With the real valued notation, the variable vector of the optimization problem, with length $2M+1$, can be denoted by $\boldsymbol{v} = [\epsilon, \boldsymbol{x}_{\mathrm{r}}^T]^T$. With this, the real valued problem reads as

$$\begin{aligned} \boldsymbol{v}_{\mathrm{opt}} = &\arg\min_{\boldsymbol{v}} \boldsymbol{a}^T\boldsymbol{v} \\ &\text{s.t. } \boldsymbol{A}\boldsymbol{v} \leq \boldsymbol{0}_{2K}, \\ &\{\boldsymbol{v}_{2m} + j\boldsymbol{v}_{2m+1}\} \in \mathcal{X}, \quad \text{for } m = 1, \ldots, M, \end{aligned} \tag{4-11}$$

with

$$\boldsymbol{a} = [-1, \boldsymbol{0}_{2M}^T]^T, \quad \boldsymbol{A} = \left[\boldsymbol{1}_{2K}, \boldsymbol{H}_{\mathrm{r}}\right].$$

Replacing the discrete solution set by its convex hull yields the relaxed problem given by

$$\boldsymbol{v}_{\mathrm{lb}} = \arg\min_{\boldsymbol{v}} \boldsymbol{a}^T\boldsymbol{v} \quad \text{s.t. } \boldsymbol{U}\boldsymbol{v} \leq \boldsymbol{p}, \tag{4-12}$$

with

$$\boldsymbol{U} = \left[\boldsymbol{A}^T, \boldsymbol{R}^T\right]^T \quad \boldsymbol{R} = \left[\boldsymbol{0}_{M\alpha_x}, \boldsymbol{R}'\right]$$

$$\boldsymbol{R}' = \left[(\boldsymbol{I}_M \otimes \boldsymbol{\beta}_1)^T, (\boldsymbol{I}_M \otimes \boldsymbol{\beta}_2)^T, \ldots, (\boldsymbol{I}_M \otimes \boldsymbol{\beta}_{\alpha_x})^T\right]^T$$

$$\boldsymbol{\beta}_i = \left[\cos\phi_i, -\sin\phi_i\right] \quad \boldsymbol{p} = \left[\boldsymbol{0}_{2K}, \tfrac{\cos(\frac{\pi}{\alpha_x})}{\sqrt{M}}\boldsymbol{1}_{M\alpha_x}\right]^T.$$

In the branch-and-bound method in each visited node subproblems are solved due to $\boldsymbol{v} = \left[\epsilon, \boldsymbol{x}_{r_1}^T, \boldsymbol{x}_{r_2}^T\right]^T$, where $\boldsymbol{x}_{r_1}$ is a fixed vector of length $2d$, which belongs to the discrete set according to $\boldsymbol{v}_{1_{2m}} + j\boldsymbol{v}_{1_{2m+1}} \in \mathcal{X}$, for $m = 1, \ldots, d$.

The matrix $\boldsymbol{U}$ can be expressed with the following structure $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{U}_1, \boldsymbol{U}_2]$, where $\boldsymbol{U}_1$ contains $2d$ columns of $\boldsymbol{U}$ and $\boldsymbol{u}_1$ is the first column of $\boldsymbol{U}$. With this, the matrix $\tilde{\boldsymbol{U}} = \left[\boldsymbol{u}_1, \boldsymbol{U}_2\right]$ and the vector $\tilde{\boldsymbol{v}} = \left[\epsilon, \boldsymbol{x}_{r_2}^T\right]^T$ are composed. Using $\tilde{\boldsymbol{U}}$ and $\tilde{\boldsymbol{v}}$ the subproblem for the lower-bounding step can be expressed as

$$\tilde{\boldsymbol{v}}_{\text{lb}} = \arg\min_{\tilde{\boldsymbol{v}}} \tilde{\boldsymbol{a}}^T\tilde{\boldsymbol{v}} \quad \text{s.t. } \tilde{\boldsymbol{U}}\tilde{\boldsymbol{v}} \leq \boldsymbol{b}, \tag{4-13}$$

with $\tilde{\boldsymbol{a}} = \left[-1, \boldsymbol{0}_{2M-2d}^T\right]^T$ and $\boldsymbol{b} = \boldsymbol{p} - \boldsymbol{U}_1\boldsymbol{x}_{r_1}$. Solving (4-13) provides a lower bound on the optimal value of the discrete problem with the condition on $\boldsymbol{x}_{r_1}$. In case the lower bound conditioned on $\boldsymbol{x}_{r_1}$ is higher than any upper bound on the original problem $\boldsymbol{x}_{r_1}$ cannot be part of the solution and every member of the discrete solution set which includes $\boldsymbol{x}_{r_1}$ can be excluded from the search. In the context of the tree search it means that once a partial solution $\boldsymbol{x}_{r_1}$ is excluded the corresponding node and all its evolutions can be skipped. The steps of the method are detailed in Algorithm 2.

Algorithm 2 Proposed B&B Precoding for solving (4-7)

initialization:

Given the channel $\boldsymbol{H}$ and transmit symbols $\boldsymbol{s}$ compute a valid upper bound $\check{g}$ on the problem in (4-7), e.g., by solving (4-8) followed by a mapping to the closest precoding vector $\boldsymbol{x} \in \mathcal{X}^M$

Define the first level ($d = 1$) of the tree by $\mathcal{G}_d := \mathcal{X}$

for $d = 1 : M - 1$ **do**

Partition $\mathcal{G}_d$ in $\boldsymbol{x}_{1,1}, \ldots, \boldsymbol{x}_{1,|\mathcal{G}_d|}$

for $i = 1 : |\mathcal{G}_d|$ **do**

Express $\boldsymbol{x}_{1,i}$ with stacked vector notation due to (4-9) as $\boldsymbol{x}_{\boldsymbol{r}1,i}$
Conditioned on $\boldsymbol{x}_{\boldsymbol{r}1,i}$ solve $\tilde{\boldsymbol{v}}_{\text{lb}}$ from (4-13)
Determine $\epsilon = [\tilde{\boldsymbol{v}}_{\text{lb}}]_1$
Compute the lower bound: $\text{lb}(\boldsymbol{x}_{1,i}) := -\epsilon$;

Map $\boldsymbol{x}_{2,\text{lb}}$ to the discrete solution with the closest
Euclidean distance: $\check{\boldsymbol{x}}_2(\boldsymbol{x}_{2,\text{lb}}) \in \mathcal{X}^{M-d}$
Using $\check{\boldsymbol{x}}_2$ find the smallest (negative) distance to the
decision threshold $\text{ub}(\boldsymbol{x}_{1,i}) :=$

$$\max_k \left[\left| \text{Im}\left\{ \boldsymbol{H}_{s^*} \begin{bmatrix} \boldsymbol{x}_{1,i} \\ \check{\boldsymbol{x}}_2 \end{bmatrix} \right\} \right| \cos\theta \} - \text{Re}\left\{ \boldsymbol{H}_{s^*} \begin{bmatrix} \boldsymbol{x}_{1,i} \\ \check{\boldsymbol{x}}_2 \end{bmatrix} \right\} \sin\theta \right]_k$$

Update the best upper bound with:
$\check{g} = \min\left(\check{g}, \text{ub}(\boldsymbol{x}_{1,i})\right)$

end for

Build a reduced set by comparing conditioned
lower bounds with the global upper bound $\check{g}$
$\mathcal{G}'_d := \{\boldsymbol{x}_{2,i} | \text{lb}(\boldsymbol{x}_{2,i}) \leq \check{g}, i = 1, \ldots, |\mathcal{G}_d|\}$

Define the set for the next level in the tree $\mathcal{G}_{d+1} := \mathcal{G}'_d \times \mathcal{X}$

end for

Search method for the ultimate level $d = M$,
Partition $\mathcal{G}_M$ in $\boldsymbol{x}_{1,1}, \ldots, \boldsymbol{x}_{1,|\mathcal{G}_M|}$

$$\epsilon(\boldsymbol{x}_{1,i}) := \min_k \left[\text{Re}\left\{ \boldsymbol{H}_{s^*} \boldsymbol{x}_{1,i} \right\} \sin\theta - \left| \text{Im}\left\{ \boldsymbol{H}_{s^*} \boldsymbol{x}_{1,i} \right\} \right| \cos\theta \right]_k$$

The global solution is

$$\boldsymbol{x}_{\text{opt}} = \arg\max_{\boldsymbol{x}_{1,i} \in \mathcal{G}_M} \epsilon(\boldsymbol{x}_{1,i})$$

4.3 Proposed Optimal and Suboptimal MMSE Precoder Design

This section exposes the objective of the precoding algorithm, discuss the chosen criterion and propose two new algorithms for low resolution precoding with phase quantization.

Unlike the state-of-the-art discrete precoding algorithm for the MMSE criterion [26], which is devised with 1-bit resolution and uses the PAPC for the bounding steps, the present study implies arbitrary uniform phase quantization and the consideration of polyhedral constraints like the ones presented in the previous section and similar to [20].

The bounding method based on polyhedra is more promising than the PAPC formulation, because the corresponding set corresponds to the convex hull and is per definition the smallest convex set that includes all the discrete solutions. Consequently, the corresponding lower bounds are more restrictive, such that the bounds can only be larger or equal to the strategy proposed in [26], which is beneficial for reducing candidates when applying a branch-and-bound method and also for finding suboptimal solutions.

4.3.1 The Continuous Problem

Using the TPC the MMSE problem can be cast as

$$\begin{aligned} &\min_{\boldsymbol{x},f} \mathrm{E}\{\|f\boldsymbol{z}-\boldsymbol{s}\|_2^2\} \\ &\text{subject to: } \boldsymbol{x}^H\boldsymbol{x} \leq E_{\mathrm{tx}}, \quad f>0. \end{aligned} \tag{4-14}$$

One approach to solve (4-14) in closed form is based on KKT conditions and the consideration that, for the optimal precoding vector, the transmit energy constraint must hold with equality as described in [13]. Then the optimal precoding vector reads as

$$\boldsymbol{x} = f^{-1}\left(\boldsymbol{H}^H\boldsymbol{H} + \frac{\mathrm{E}\left\{\boldsymbol{w}^H\boldsymbol{w}\right\}}{E_{\mathrm{tx}}}\boldsymbol{I}\right)^{-1}\boldsymbol{H}^H\boldsymbol{s}, \tag{4-15}$$

where the optimal scaling factor is given by

$$f = \sqrt{\frac{1}{E_{\mathrm{tx}}}\left(\boldsymbol{s}^H\boldsymbol{H}\left(\boldsymbol{H}^H\boldsymbol{H} + \frac{\mathrm{E}\{\boldsymbol{w}^H\boldsymbol{w}\}}{E_{\mathrm{tx}}}\boldsymbol{I}\right)^{-2}\boldsymbol{H}^H\boldsymbol{s}\right)}. \tag{4-16}$$

4.3.2 Problem for Constant Envelope Signals With Phase Quantization at the Transmitter and PSK Modulation

The problem desribed in (4-14) considers infinite resolution for the entries in $\boldsymbol{x}$. Considering quantization of the transmit signal yields the restriction to a discrete input alphabet such that the corresponding problem can be cast as

$$\begin{aligned} &\min_{\boldsymbol{x},f} \mathrm{E}\{\|f\boldsymbol{z}-\boldsymbol{s}\|_2^2\} \\ &\text{subject to: } \boldsymbol{x}\in\mathcal{X}^M, \quad f>0. \end{aligned} \tag{4-17}$$

Note that the feasible set is discrete and therefore not convex. In addition, it can be shown that the MMSE objective function in (4-17) is not jointly convex in $\boldsymbol{x}$ and f as described in the Appendix A.1. Accordingly the optimization problem is not convex.

In the following, a suboptimal algorithm based on the relaxation of the feasible set and formulation of an equivalent convex problem is devised. Subsequently, a branch-and-bound strategy for computing the optimal precoding vector is formulated.

4.3.2.1 Proposed MMSE Mapped Precoder

In this subsection, a suboptimal approach for the problem described in (4-17) is proposed. Since the feasible set, $\mathcal{X}^M$, of the optimization problem presented by (4-17) is non-convex it is replaced by its convex hull $\mathcal{P}$, which is a polyhedron. With this, the problem reads as

$$\begin{aligned} &\min_{\boldsymbol{x},f} \mathrm{E}\{\|f(\boldsymbol{H}\boldsymbol{x}+\boldsymbol{w})-\boldsymbol{s}\|_2^2\} \\ &\text{subject to: } \boldsymbol{x}\in\mathcal{P}, \quad f>0. \end{aligned} \tag{4-18}$$

Rewriting the problem in a real valued notation yields

$$\begin{aligned} &\min_{\boldsymbol{x}_\mathrm{r},f} \mathrm{E}\{\|f(\boldsymbol{H}_\mathrm{r}\ \boldsymbol{x}_\mathrm{r}+\boldsymbol{w}_\mathrm{r})-\boldsymbol{s}_\mathrm{r}\|_2^2\} \\ &\text{subject to: } \boldsymbol{A}\boldsymbol{x}_\mathrm{r}\leq\boldsymbol{b}, \quad f>0, \end{aligned} \tag{4-19}$$

with

$$\begin{aligned}\boldsymbol{x}_{\mathrm{r}} &= \begin{bmatrix}\mathrm{Re}\{\boldsymbol{x}_1\} & \mathrm{Im}\{\boldsymbol{x}_1\} & \cdots & \mathrm{Re}\{\boldsymbol{x}_M\} & \mathrm{Im}\{\boldsymbol{x}_M\}\end{bmatrix}^T,\\ \boldsymbol{w}_{\mathrm{r}} &= \begin{bmatrix}\mathrm{Re}\{\boldsymbol{w}_1\} & \mathrm{Im}\{\boldsymbol{w}_1\} & \cdots & \mathrm{Re}\{\boldsymbol{w}_K\} & \mathrm{Im}\{\boldsymbol{w}_K\}\end{bmatrix}^T,\\ \boldsymbol{s}_{\mathrm{r}} &= \begin{bmatrix}\mathrm{Re}\{\boldsymbol{s}_1\} & \mathrm{Im}\{\boldsymbol{s}_1\} & \cdots & \mathrm{Re}\{\boldsymbol{s}_K\} & \mathrm{Im}\{\boldsymbol{s}_K\}\end{bmatrix}^T\end{aligned}$$

and

$$\boldsymbol{H}_{\mathrm{r}} = \begin{bmatrix}\mathrm{Re}\{h_{11}\} & -\mathrm{Im}\{h_{11}\} & \cdots & \mathrm{Re}\{h_{1M}\} & -\mathrm{Im}\{h_{1M}\}\\ \mathrm{Im}\{h_{11}\} & \mathrm{Re}\{h_{11}\} & \cdots & \mathrm{Im}\{h_{1M}\} & \mathrm{Re}\{h_{1M}\}\\ \vdots & & \ddots & & \vdots\\ \mathrm{Re}\{h_{K1}\} & -\mathrm{Im}\{h_{K1}\} & \cdots & \mathrm{Re}\{h_{KM}\} & -\mathrm{Im}\{h_{KM}\}\\ \mathrm{Im}\{h_{K1}\} & \mathrm{Re}\{h_{K1}\} & \cdots & \mathrm{Im}\{h_{KM}\} & \mathrm{Re}\{h_{KM}\}\end{bmatrix}.$$

The inequality $\boldsymbol{A}\boldsymbol{x}_{\mathrm{r}} \leq \boldsymbol{b}$ restricts the elements of the precoding vector to be inside or on the boarder of the polyhedron whose construction will be detailed in the sequel. An equivalent problem to (4-18) can be cast as

$$\begin{aligned}&\min_{\boldsymbol{x}_{\mathrm{r}},f} f^2\boldsymbol{x}_{\mathrm{r}}^T\boldsymbol{H}_{\mathrm{r}}^T\boldsymbol{H}_{\mathrm{r}}\,\boldsymbol{x}_{\mathrm{r}} - 2f\boldsymbol{x}_{\mathrm{r}}^T\boldsymbol{H}_{\mathrm{r}}^T\boldsymbol{s}_{\mathrm{r}} + f^2\mathrm{E}\{\boldsymbol{w}_{\mathrm{r}}^T\boldsymbol{w}_{\mathrm{r}}\} \qquad (4\text{-}20)\\ &\text{subject to: } \boldsymbol{A}\boldsymbol{x}_{\mathrm{r}} \leq \boldsymbol{b}, \quad f > 0.\end{aligned}$$

If $f > 0$ would be constant, the problem would be a convex quadratic program, since $\boldsymbol{H}_{\mathrm{r}}^T\boldsymbol{H}_{\mathrm{r}} \in S_+^n$ (cf. Section 4.4 in [42]). Yet, as mentioned before, the problem is in general not jointly convex in f and $\boldsymbol{x}_{\mathrm{r}}$, as can be seen in the Appendix A.1.

Nevertheless, the problem can be rewritten as an equivalent convex problem by shifting the scaling factor f to the constraints and substituting the optimization variable. This essentially means that the feasible set is scaled depending on the value of f. In this context, the mentioned substitution is applied and by introducing a new optimization variable $\boldsymbol{x}_{\mathrm{r,f}} = f\ \boldsymbol{x}_{\mathrm{r}}$. Accordingly, the resulting problem reads as

$$\begin{aligned}&\min_{\boldsymbol{x}_{\mathrm{r,f}},f} \boldsymbol{x}_{\mathrm{r,f}}^T\boldsymbol{H}_{\mathrm{r}}^T\boldsymbol{H}_{\mathrm{r}}\boldsymbol{x}_{\mathrm{r,f}} - 2\boldsymbol{x}_{\mathrm{r,f}}^T\boldsymbol{H}_{\mathrm{r}}^T\boldsymbol{s}_{\mathrm{r}} + f^2\mathrm{E}\{\boldsymbol{w}_{\mathrm{r}}^T\boldsymbol{w}_{\mathrm{r}}\} \qquad (4\text{-}21)\\ &\text{subject to: } \boldsymbol{A}\boldsymbol{x}_{\mathrm{r,f}} \leq f\boldsymbol{b}, \qquad f > 0.\end{aligned}$$

The first constraint can be rewritten as a linear constraint, such that the

problem is a convex quadratic program (cf. Section 4.4 in [42]), which reads as

$$\min_{\boldsymbol{x}_{\mathrm{r,f}},f} \boldsymbol{x}_{\mathrm{r,f}}^T \boldsymbol{H}_{\mathrm{r}}^T \boldsymbol{H}_{\mathrm{r}} \boldsymbol{x}_{\mathrm{r,f}} - 2\boldsymbol{x}_{\mathrm{r,f}}^T \boldsymbol{H}_{\mathrm{r}}^T \boldsymbol{s}_{\mathrm{r}} + f^2 \mathrm{E}\{\boldsymbol{w}_{\mathrm{r}}^T \boldsymbol{w}_{\mathrm{r}}\} \tag{4-22}$$

$$\text{subject to: } \boldsymbol{R} \begin{bmatrix} \boldsymbol{x}_{\mathrm{r},f} \\ f \end{bmatrix} \leq \boldsymbol{0}, \qquad f > 0,$$

where $\boldsymbol{R} = \begin{bmatrix} \boldsymbol{A} & -\boldsymbol{b} \end{bmatrix}$. The polyhedron associated to uniformly phase quantized transmit symbols with α_x different phases can be expressed as proposed in section 4.2, which is similar to the description in [20]. The corresponding matrix notation reads as

$$\boldsymbol{A} = \begin{bmatrix} (\boldsymbol{I}_M \otimes \boldsymbol{\beta}_1)^T & (\boldsymbol{I}_M \otimes \boldsymbol{\beta}_2)^T & \dots & (\boldsymbol{I}_M \otimes \boldsymbol{\beta}_{\alpha_x})^T \end{bmatrix}^T, \tag{4-23}$$

$$\boldsymbol{\beta}_i = \begin{bmatrix} \cos\phi_i & -\sin\phi_i \end{bmatrix}, \quad \phi_i = \frac{2\pi i}{\alpha_x}, \text{ for } i = 1, \dots, \alpha_x, \tag{4-24}$$

$$\boldsymbol{b} = \frac{\cos(\frac{\pi}{\alpha_x})}{\sqrt{M}} \boldsymbol{1}_{M\alpha_x}, \tag{4-25}$$

with $\boldsymbol{1}_{M\alpha_x}$ being a column vector with length $M\alpha_x$. Note that the solution of (4-22) yields a lower bound on the optimal value of the original problem, meaning that the corresponding MSE is smaller or equal to the corresponding MSE of the original problem in (4-17). Yet, the optimal solution of the relaxed problem is not necessarily in the feasible set of the original problem $\mathcal{X}^M$.

Therefore, in order to find a feasible solution, mapping to the closest Euclidean distance point in $\mathcal{X}^M$ is considered. The solution after mapping, then, yields an MSE which is always greater or equal to the optimal of (4-17), meaning that after the mapping process an upper bound on the optimal value of the original problem is found.

4.3.2.2 Proposed Optimal Approach via Branch-and-Bound

As stated before, the continuous solution of (4-22) is in general not in $\mathcal{X}^M$ and then it only provides an unfeasible lower bound, or, after mapping, a feasible upper bound solution for the original problem. In this sense, the method in (4-22) does not provide a reliable way for solving (4-17).

Therefore, a branch-and-bound strategy is proposed. This strategy always provides the optimal solution for (4-17) with significantly reduced computational complexity as compared to exhaustive search.

Branch-and-Bound Initialization

The branch-and-bound algorithm converges faster when an upper bound that permits many exclusions is computed as early as possible. Therefore, it is recommended to have an initialization step where an upper bound $\check{g} < \infty$ is found before beginning with the search process.

In this regard, for initialization, the problem described in (4-22) is solved. With this, $\boldsymbol{x}_{\mathrm{lb}}$ and $g(\boldsymbol{x}_{\mathrm{lb}}) = \mathrm{MSE}_{\mathrm{lb}}$ are obtained. After mapping $\boldsymbol{x}_{\mathrm{lb}}$ to the feasible set $\boldsymbol{x}_{\mathrm{ub}}$ and $\check{g} = \mathrm{MSE}_{\mathrm{ub}}$ are determined.

Note that if the continuous solution of (4-22) is in the feasible set, upper and lower bound are equal which can be expressed as

$$\boldsymbol{x}_{\mathrm{ub}} = \boldsymbol{x}_{\mathrm{lb}} = \boldsymbol{x}_{\mathrm{opt}} \rightarrow g(\boldsymbol{x}_{\mathrm{lb}}) = \check{g}. \tag{4-26}$$

This would mean that the optimal solution is found already by the approach from subsection 4.3.2.1 and the tree search process can be skipped.

Subproblems

When the condition from (4-26) is not met, one can apply the branch-and-bound tree search method. It is, then, necessary to solve subproblems, as first mentioned on section 4.1. The equations that define the subproblems are derived below.

First the precoding vector is divided in a fixed vector of length $2d$ and a variable vector according to

$$\boldsymbol{x}_{\mathrm{r}} = \left[\boldsymbol{x}_{\mathrm{r,fixed}}^{T} \ , \ \boldsymbol{x}_{\mathrm{r}}^{\prime T}\right]^{T}. \tag{4-27}$$

With this, the MMSE problem formulation reads as

$$\begin{aligned} &\min_{\boldsymbol{x}_{\mathrm{r}}^{\prime}, f^{\prime}} \mathrm{E}\{\| f^{\prime}(\boldsymbol{H}_{\mathrm{r}} \left[\boldsymbol{x}_{\mathrm{r,fixed}}^{T} \ , \ \boldsymbol{x}_{\mathrm{r}}^{\prime T}\right]^{T} + \boldsymbol{w}_{\mathrm{r}}) - \boldsymbol{s}_{\mathrm{r}} \|_2^2\} \\ &\text{subject to: } \boldsymbol{A}^{\prime} \boldsymbol{x}_{\mathrm{r}}^{\prime} \leq \boldsymbol{b}^{\prime}, \quad f^{\prime} > 0, \end{aligned} \tag{4-28}$$

where $\boldsymbol{A}^{\prime} \boldsymbol{x}_{\mathrm{r}}^{\prime} \leq \boldsymbol{b}^{\prime}$ restricts the elements of the precoding vector to be inside of the set $\mathcal{J}$ and will be detailed in what follows. The channel can be rewritten accordingly as $\boldsymbol{H}_{\mathrm{r}} = [\boldsymbol{H}_{\mathrm{r,\ fixed}} \ \ \boldsymbol{H}_{\mathrm{r}}^{\prime}]$. Then the problem can be cast as

$$\begin{aligned} &\min_{\boldsymbol{x}_{\mathrm{r}}^{\prime}, f^{\prime}} \mathrm{E}\{\| f^{\prime}(\boldsymbol{H}_{\mathrm{r}}^{\prime} \ \boldsymbol{x}_{\mathrm{r}}^{\prime} + \boldsymbol{H}_{\mathrm{r,\ fixed}} \ \boldsymbol{x}_{\mathrm{r,\ fixed}} + \boldsymbol{w}_{\mathrm{r}}) - \boldsymbol{s}_{\mathrm{r}} \|_2^2\} \\ &\text{subject to: } \boldsymbol{A}^{\prime} \boldsymbol{x}_{\mathrm{r}}^{\prime} \leq \boldsymbol{b}^{\prime}, \quad f^{\prime} > 0, \end{aligned} \tag{4-29}$$

and an equivalent problem is given by

$$\begin{aligned} &\min_{\boldsymbol{x}'_{\mathrm{r}},f'} \|f'\boldsymbol{H}'_{\mathrm{r}}\,\boldsymbol{x}'_{\mathrm{r}} - \boldsymbol{s}_{\mathrm{r}} + f'\boldsymbol{H}_{\mathrm{r,\,fixed}}\,\boldsymbol{x}_{\mathrm{r,\,fixed}}\|_2^2 + f'^2\mathrm{E}\{\boldsymbol{w}_{\mathrm{r}}^T\boldsymbol{w}_{\mathrm{r}}\} \\ &\text{subject to: } \boldsymbol{A}'\boldsymbol{x}'_{\mathrm{r}} \leq \boldsymbol{b}', \quad f' > 0. \end{aligned} \tag{4-30}$$

The objective function is not jointly convex in f' and $\boldsymbol{x}'_{\mathrm{r}}$, as shown for the conventional MMSE cost function in the Appendix A.1.

However, it is possible to shift the scaling factor from the objective in the polyhedron constraint as done in subsection 4.3.2.1. Accordingly, the variable $\boldsymbol{x}'_{\mathrm{r},f}$ is introduced as $f'\boldsymbol{x}'_{\mathrm{r}} = \boldsymbol{x}'_{\mathrm{r},f}$. Using $\boldsymbol{x}'_{\mathrm{r},f}$, the equivalent problem reads as

$$\begin{aligned} &\min_{\boldsymbol{x}'_{\mathrm{r},f},f'} \|\boldsymbol{H}'_{\mathrm{r}}\,\boldsymbol{x}'_{\mathrm{r},f} - \boldsymbol{s}_{\mathrm{r}} + f'\boldsymbol{H}_{\mathrm{r,\,fixed}}\,\boldsymbol{x}_{\mathrm{r,\,fixed}}\|_2^2 + f'^2\mathrm{E}\{\boldsymbol{w}_{\mathrm{r}}^T\boldsymbol{w}_{\mathrm{r}}\} \\ &\text{subject to: } \boldsymbol{A}'\boldsymbol{x}'_{\mathrm{r},f} \leq f'\boldsymbol{b}', \quad f' > 0. \end{aligned} \tag{4-31}$$

Rearranging the constraint yields

$$\begin{aligned} &\min_{\boldsymbol{x}'_{\mathrm{r},f},f'} \|\boldsymbol{H}'_{\mathrm{r}}\,\boldsymbol{x}'_{\mathrm{r},f} - \boldsymbol{s}_{\mathrm{r}} + f'\boldsymbol{H}_{\mathrm{r,\,fixed}}\,\boldsymbol{x}_{\mathrm{r,\,fixed}}\|_2^2 + f'^2\mathrm{E}\{\boldsymbol{w}_{\mathrm{r}}^T\boldsymbol{w}_{\mathrm{r}}\} \\ &\text{subject to: } \boldsymbol{R}' \begin{bmatrix} \boldsymbol{x}'_{\mathrm{r},f} \\ f' \end{bmatrix} \leq \boldsymbol{0}, \quad f' > 0, \end{aligned} \tag{4-32}$$

where $\boldsymbol{R}' = [\boldsymbol{A}' - \boldsymbol{b}']$ is obtained by selecting the last $2\,(M-d)$ columns of $\boldsymbol{R}$.

Note that the problem in (4-32) is convex because of the convex constraints and its objective function which is jointly convex in f' and $\boldsymbol{x}'_{\mathrm{r},f}$ as can be seen in the Appendix A.2, where the Hessian is examined.

MMSE Branch-and-Bound Precoding Algorithm

In this subsection, a branch-and-bound algorithm is proposed which solves (4-17) with the tools presented in the previous subsections. As mentioned before, the first step is the initialization, where the problem from (4-22) is solved and the condition $\boldsymbol{x}_{\mathrm{lb}} = \boldsymbol{x}_{\mathrm{ub}}$ is evaluated. If the condition is met, the algorithm returns $\boldsymbol{x}_{\mathrm{lb}}$. Otherwise, the branch-and-bound tree search process is performed as described in the sequel.

For the tree search process a breadth first search is devised and the subproblems are solved considering partially fixed precoding vectors $\boldsymbol{x}_{\mathrm{r}} = \left[\boldsymbol{x}_{\mathrm{r,fixed}}^T \;\; \boldsymbol{x}_{\mathrm{r}}'^T\right]^T$, where $\boldsymbol{x}_{\mathrm{r,fixed}}$ has length $2d$ as previously stated. Accordingly, the matrix $\boldsymbol{H}_{\mathrm{r}}$ is divided as $\boldsymbol{H}_{\mathrm{r}} = [\boldsymbol{H}_{\mathrm{r,fixed}} \;\; \boldsymbol{H}'_{\mathrm{r}}]$, where $\boldsymbol{H}_{\mathrm{r,fixed}}$ contains the first $2d$ columns of $\boldsymbol{H}_{\mathrm{r}}$. Moreover the matrix $\boldsymbol{R}'$ is obtained via selecting the

last $2\,(M-d)$ columns of $\boldsymbol{R}$.

Using $\boldsymbol{R}'$ and $\boldsymbol{H}'_{\mathrm{r}}$, the subproblem (4-32) for the lower-bounding step is solved. Mapping the solution from (4-32) to the discrete set yields $\boldsymbol{x}_{\mathrm{r,ub}}$. Based on $\boldsymbol{x}_{\mathrm{r,ub}}$, the MSE is minimized with choosing

$$f' = \frac{\boldsymbol{s}_{\mathrm{r}}^T \boldsymbol{H}_{\mathrm{r}} \begin{bmatrix} \boldsymbol{x}_{\mathrm{r,fixed}} \\ \boldsymbol{x}_{\mathrm{r,ub}} \end{bmatrix}}{\left\| \boldsymbol{H}_{\mathrm{r}} \begin{bmatrix} \boldsymbol{x}_{\mathrm{r,fixed}} \\ \boldsymbol{x}_{\mathrm{r,ub}} \end{bmatrix} \right\|_2^2 + \mathrm{E}\{\boldsymbol{w}_{\mathrm{r}}^T \boldsymbol{w}_{\mathrm{r}}\}}. \tag{4-33}$$

The corresponding MSE serves as an upper bound on the optimal value of the original problem ($\mathrm{MSE}_{\mathrm{ub}}$).

In case the lower bound conditioned on $\boldsymbol{x}_{\mathrm{r,fixed}}$ is higher than any upper bound on the original problem, $\boldsymbol{x}_{\mathrm{r,fixed}}$ cannot be part of the solution and every member of the discrete solution set which includes $\boldsymbol{x}_{\mathrm{r,fixed}}$ can be excluded from the search process. The steps of the method are detailed in Algorithm 3.

Note that, when operating in the high-SNR regime, the computation of the optimal precoding vector in each symbol period can correspond to enormous computational complexity. The precomputation of the lookup-table $\mathcal{L}$ is especially important in this situation, since it allows the precoding method to be a practical solution for channels with large coherence time, as suggested in [43].

Algorithm 3 Proposed B&B Precoding for solving (4-17)

initialization:

Given the channel $\boldsymbol{H}$ and transmit symbols $\boldsymbol{s}$ compute a valid upper bound $\check{g}$ on the problem in (4-17), by solving (4-22) followed by a mapping to the closest precoding vector $\boldsymbol{x} \in \mathcal{X}^M$ and computing its MSE. If the solution of (4-22) belongs to $\mathcal{X}^M$ it is optimal. Otherwise, define the first level ($d = 1$) of the tree by $\mathcal{G}_d := \mathcal{X}$

for $d = 1 : M - 1$ **do**

Partition $\mathcal{G}_d$ in $\boldsymbol{x}_{\text{fixed},1}, \ldots, \boldsymbol{x}_{\text{fixed},|\mathcal{G}_d|}$

for $i = 1 : |\mathcal{G}_d|$ **do**

Express $\boldsymbol{x}_{\text{fixed},i}$ in real valued notation $\boldsymbol{x}_{\text{r,fixed},i}$

Conditioned on $\boldsymbol{x}_{\text{r,fixed},i}$ solve (4-32) to find $\boldsymbol{x}'_{\text{r},f}$ and f'

Determine the lower bound as:

$$\text{MSE}_{\text{lb}} := \left|\left|\boldsymbol{H}'_{\text{r}}\ \boldsymbol{x}'_{\text{r},f} - \boldsymbol{s}_{\text{r}} + f'\boldsymbol{H}_{\text{r, fixed}}\ \boldsymbol{x}_{\text{r, fixed,i}}\right|\right|_2^2 + f'^2\text{E}\{\boldsymbol{w}_{\text{r}}^T\boldsymbol{w}_{\text{r}}\}$$

Extract $\boldsymbol{x}'_{\text{r}} = \frac{\boldsymbol{x}'_{r,f}}{f'}$

Rewrite $\boldsymbol{x}'_{\text{r}}$ in complex notation as $\boldsymbol{x}'_{\text{lb}}$

Map $\boldsymbol{x}'_{\text{lb}}$ to the discrete solution with the closest Euclidean distance: $\boldsymbol{x}'_{\text{ub}}(\boldsymbol{x}'_{\text{lb}}) \in \mathcal{X}^{M-d}$

Express $\boldsymbol{x}'_{\text{ub}}$ in real valued notation $\boldsymbol{x}'_{\text{r,ub}}$

Compute f' according to (4-33)

With $\boldsymbol{x}'_{\text{r,ub}}$ and f', the upper bound is

$$\text{MSE}_{\text{ub}}(\boldsymbol{x}_{\text{ r,fixed,i}}) := \left|\left|\ f'\boldsymbol{H}_{\text{r}}\left[\boldsymbol{x}_{\text{r,fixed,i}}\ \ \boldsymbol{x}'_{\text{r,ub}}\right] - \boldsymbol{s}_{\text{r}}\right|\right|_2^2 + f'^2\text{E}\{\boldsymbol{w}_{\text{r}}^T\boldsymbol{w}_{\text{r}}\}$$

Update the best upper bound with $\check{g} = \min(\check{g}, \text{MSE}_{\text{ub}})$

end for

Construct a reduced set by comparing conditioned lower bounds with the global upper bound $\check{g}$

$\mathcal{G}'_d := \left\{\boldsymbol{x}'_{\text{lb},i} | \text{MSE}_{\text{lb}}(\boldsymbol{x}'_{\text{lb},i}) \leq \check{g}, i = 1, \ldots, |\mathcal{G}_d|\right\}$

Define the set for the next level in the tree: $\mathcal{G}_{d+1} := \mathcal{G}'_d \times \mathcal{X}$

end for

Search method for the ultimate level $d = M$,

Partition $\mathcal{G}_M$ in $\boldsymbol{x}_{\text{fixed},1}, \ldots, \boldsymbol{x}_{\text{fixed},|\mathcal{G}_M|}$

Express $\boldsymbol{x}_{\text{fixed},i}$ with real valued notation $\boldsymbol{x}_{\text{r,fixed},i}$ and compute f' with (4-33)

$$\text{MSE}(\boldsymbol{x}_{\text{fixed},i}) := ||f'\boldsymbol{H}_{\text{r}}\boldsymbol{x}_{\text{r,fixed,i}} - \boldsymbol{s}_{\text{r}}||_2^2 + f'^2\text{E}\{\boldsymbol{w}_{\text{r}}^T\boldsymbol{w}_{\text{r}}\}$$

The global solution is :

$$\boldsymbol{x}_{\text{opt}} = \arg\min_{\boldsymbol{x}_{\text{fixed},i} \in \mathcal{G}_M} \text{MSE}(\boldsymbol{x}_{\text{fixed},i})$$

4.4 Final Considerations

4.4.1 On Optimality

The proposed branch-and-bound algorithms are optimal in the sense of providing the precoding vector that yields the maximum MDDT and the minimum MSE. However, the presented algorithms are not necessarily optimal in the sense of BER. The MMDDT and MMSE criteria were chosen for the development of the proposed precoders once the minimization of the BER (Min BER) criterion is difficult to achieve for all values of SNR.

4.4.2 On the Computational Complexity

The computational complexity of both proposed branch-and-bound approaches, as will be seen in subsection 6.1.2, grows with the number of BS antennas. In the case of the MMSE branch-and-bound algorithm, it also increases with the SNR.

In massive multiuser MIMO systems, the computation of the optimal precoding vector in each symbol period can correspond to enormous computational complexity. In the case of using the proposed MMSE branch-and-bound method, this happens, particularly in the high SNR regime. The precomputation of the lookup-table $\mathcal{L}$ is especially important in this situation since it allows the proposed precoding methods to be a practical solution for channels with large coherence time, as suggested in [43].

4.4.3 On the Computation of the Lookup-Table

For forming $\mathcal{L}$ it is not always necessary to compute $\boldsymbol{x}(\boldsymbol{s}) \ \forall \ \boldsymbol{s} \in \mathcal{S}^K$. If the utilized precoding strategy has the circular symmetric property, exposed by the following equation,

$$\phi = \frac{\pi(2i+1)}{\alpha_x}, \ i = 1...\alpha_x \implies \boldsymbol{x}\left(\boldsymbol{s}\, e^{\,j\phi}\right) = \boldsymbol{x}(\boldsymbol{s})\, e^{\,j\phi}, \tag{4-34}$$

when $\alpha_x = \alpha_s$, $\boldsymbol{x}(\boldsymbol{s})$ can be computed by applying a rotation to another previously computed precoding vector. In this case, the number of precoding vectors precomputed is reduced which decreases the hardware costs necessary for its implementation.

5
Discrete Precoding Aware Receiver Design

This section exposes the design of the DPA-IDD Receiver where three soft detection methods for the computation of the extrinsic information are proposed. The objective is to enable channel coding in conjunction with discrete precoding methods, as those proposed in chapter 4.

The section is divided into two parts, the first proposes three soft detection approaches for computing the extrinsic information, while the second describes the DPA-IDD scheme.

The objective of the DPA-IDD receiver is to compute LLRs that are used to make a decision about $\boldsymbol{c}_k$ which implies $\boldsymbol{m}_k$. The LLRs are defined as follows

$$L(c_{k,i}) = \ln\left(\frac{P\left(c_{k,i}=+1|z_k[t]\right)}{P\left(c_{k,i}=-1|z_k[t]\right)}\right), \tag{5-1}$$

where $z_k[t]$ is the received signal and $c_{k,i} \in \{-1,+1\}$. Using Bayes' theorem, equation (5-1) is rewritten as

$$\begin{aligned} L(c_{k,i}) &= \ln\left(\frac{p\left(z_k[t]|c_{k,i}=+1\right)}{p\left(z_k[t]|c_{k,i}=-1\right)}\right) + \ln\left(\frac{P\left(c_{k,i}=+1\right)}{P\left(c_{k,i}=-1\right)}\right) \\ &= L_e\left(c_{k,i}\right) + L_a\left(c_{k,i}\right), \end{aligned} \tag{5-2}$$

where $L_e\left(c_{k,i}\right)$ and $L_a\left(c_{k,i}\right)$ denote the extrinsic and a priori information, respectively.

5.1
Extrinsic Information Computation

In this section, three methods for computing $L_e(c_{k,i})$ are presented. The first method computes the extrinsic information based on the true PDF of the received signal, while the second relies on a nonlinear Gaussian approximation of the original PDF. The third approach calculates $L_e(c_{k,i})$ by relying on a linear model. As shown in equation (5-2), the $L_e(c_{k,i})$ is defined as

$$L_e\left(c_{k,i}\right) = \ln\left(\frac{p\left(z_k[t]|c_{k,i}=+1\right)}{p\left(z_k[t]|c_{k,i}=-1\right)}\right). \tag{5-3}$$

Using the law of total probability equation (5-3) can be expanded as

$$L_e\left(c_{k,i}\right) = \ln\left(\frac{\sum\limits_{s\in S_{+1}} p\left(z_k[t]|s\right) P\left(s|r_{k,t,\upsilon}=+1\right)}{\sum\limits_{s\in S_{-1}} p\left(z_k[t]|s\right) P\left(s|r_{k,t,\upsilon}=-1\right)}\right), \tag{5-4}$$

with $\upsilon = (i-(t-1)N) \in \{1,\ldots,N\}$. The sets S_{+1} and S_{-1} represent all possible constellation points where the υ-th bit of $\boldsymbol{r}_{k,t}$ is $+1$ or -1, respectively. For a given $s \in S_g$, $g \in \{+1,-1\}$, if $\mathcal{M}^{-1}(s) = [a_1,\ldots,a_\upsilon = g,\ldots,a_N]$, the probability $P\left(s|r_{k,t,\upsilon}=g\right)$ is given by

$$P\left(s|r_{k,t,\upsilon}=g\right) = \prod_{\substack{l=1\\ l\neq\upsilon}}^{N} P(r_{k,t,l}=a_l). \tag{5-5}$$

Based on the a priori information, the previous equation is rewritten as

$$P\left(s|r_{k,t,\upsilon}=g\right) = \prod_{\substack{l=1\\ l\neq\upsilon}}^{N} \frac{e^{\left(a_l\ L_a(r_{k,t,l})\right)}}{1+e^{\left(a_l\ L_a(r_{k,t,l})\right)}}, \tag{5-6}$$

where $L_a\left(r_{k,t,l}\right) = L_a\left(c_{k,l+(t-1)N}\right)$. Note that for computing $L_e(c_{k,i})$ the only demands are $p\left(z_k[t]|s\right)$, that need to be known for all $s \in \mathcal{S}$, and the knowledge of $L_a\left(c_{k,i}\right)$ for $i = 1,\ldots,\frac{N_b}{R}$.

5.1.1 Common AWGN Approach

A common approach for the computation of $L_e(c_{k,i})$ relies on the assumption of an AWGN channel. It then considers that the received signal can be described as shown in the following

$$z_k[t] = s_k[t] + w_k[t]. \tag{5-7}$$

With this, the PDF of the received signal is a complex random Gaussian, meaning $z_k\left[t\right] \sim \mathcal{CN}(s_k[t],\sigma_w^2)$. Finally, $p(z_k[t]|s)$ is given by

$$p\left(z_k[t]|s\right) = \frac{1}{\pi\sigma_w^2}\ \mathrm{e}^{-\frac{\left|z_k[t]-s\right|^2}{\sigma_w^2}}. \tag{5-8}$$

Consequently inserting (5-8) in (5-4) yields

$$L_e\left(c_{k,i}\right) = \ln\left(\frac{\sum\limits_{s\in S_{+1}} \mathrm{e}^{-\frac{|z_k[t]-s|^2}{\sigma_w^2}}\ P\left(s|r_{k,t,v}=+1\right)}{\sum\limits_{s\in S_{-1}} \mathrm{e}^{-\frac{|z_k[t]-s|^2}{\sigma_w^2}}\ P\left(s|r_{k,t,v}=-1\right)}\right). \tag{5-9}$$

5.1.2 Discrete Precoding Aware Soft Detector

As seen in the previous subsection, the assumption of the common AWGN approach leads to a simple method for computing $L_e(c_{k,i})$. However, the consideration that the received signal can be described as in equation (5-7) is not precise for the medium and high SNR regime and yields BER performance degradation, as will be discussed in chapter 6.

In this subsection, the DPA Soft Detector, which can be considered as the soft MAP detector, is introduced as a method for computing $L_e(c_{k,i})$. First, the received signal $z_k[t]$ is rewritten in a stacked vector notation $\boldsymbol{z}_r[t] = [\mathrm{Re}\{z_k[t]\}\ \ \mathrm{Im}\{z_k[t]\}]^T$, where, for simplicity, the index k is suppressed. The distribution $p\left(z_k[t]|s\right)$ is given by

$$\begin{aligned} p\left(z_k[t]|s\right) &= \sum_{\boldsymbol{s}'\in\mathcal{S}^{K-1}} p\left(z_k[t]|s,\boldsymbol{s}'\right)\ P(\boldsymbol{s}') \\ &= \left(\frac{1}{\alpha_s}\right)^{K-1} \frac{1}{\pi\sigma_w^2} \sum_{\boldsymbol{s}'\in\mathcal{S}^{K-1}} \mathrm{e}^{-\frac{||\boldsymbol{z}_r[t]-\mathrm{E}\{\boldsymbol{z}_r[t]|s,\boldsymbol{s}'\}||_2^2}{\sigma_w^2}}, \end{aligned} \tag{5-10}$$

where $\boldsymbol{s}' = \left[s'_1,\ldots,s'_{k-1},s'_{k+1},\ldots,s'_K\right]^T$ corresponds to the symbols of the other users. For a given s and $\boldsymbol{s}'$ the expected value of the receive signal is given by

$$\mathrm{E}\left\{\boldsymbol{z}_r[t]|s,\boldsymbol{s}'\right\} = \left[\mathrm{Re}\left\{\boldsymbol{h}_k\boldsymbol{x}\left(s,\boldsymbol{s}'\right)\right\}\quad \mathrm{Im}\left\{\boldsymbol{h}_k\boldsymbol{x}\left(s,\boldsymbol{s}'\right)\right\}\right]^T. \tag{5-11}$$

With this, $L_e(c_{k,i})$ can be computed by inserting (5-10) into (5-4). The resulting expression, finally, reads as

$$L_e(c_{k,i}) = \ln\left(\frac{\sum\limits_{s\in S_{+1}}\sum\limits_{\boldsymbol{s}'\in\mathcal{S}^{K-1}} \mathrm{e}^{-\frac{||\boldsymbol{z}_r[t]-\mathrm{E}\{\boldsymbol{z}_r[t]|s,\boldsymbol{s}'\}||_2^2}{\sigma_w^2}} P\left(s|r_{k,t,v}=+1\right)}{\sum\limits_{s\in S_{-1}}\sum\limits_{\boldsymbol{s}'\in\mathcal{S}^{K-1}} \mathrm{e}^{-\frac{||\boldsymbol{z}_r[t]-\mathrm{E}\{\boldsymbol{z}_r[t]|s,\boldsymbol{s}'\}||_2^2}{\sigma_w^2}} P\left(s|r_{k,t,v}=-1\right)}\right). \tag{5-12}$$

Note that, for using (5-12), $p\left(z_k[t]|s,\boldsymbol{s}'\right)$ needs to be evaluated for all elements of $\mathcal{S}^{K-1}$. Hence, computing (5-12) can lead to a prohibitive computational

complexity at the receiver side for systems with many users.

5.1.3 Gaussian Discrete Precoding Aware Soft Detector

To reduce the computational complexity the Gaussian Discrete Precoding Aware (GDPA) Soft Detector is developed. The basic assumption is that the vector $\boldsymbol{z}_r[t]$ can be described as a Gaussian random vector, meaning

$$\tilde{p}\left(z_k[t]|s\right) = \frac{\mathrm{e}^{-\frac{1}{2}\left[\left(\boldsymbol{z}_r[t]-\boldsymbol{\mu}_{z_r|s}\right)^T \boldsymbol{C}_{z_r|s}^{-1}\left(\boldsymbol{z}_r[t]-\boldsymbol{\mu}_{z_r|s}\right)\right]}}{2\pi\sqrt{\det\left(\boldsymbol{C}_{z_r|s}\right)}}. \tag{5-13}$$

In the following the computation of $\boldsymbol{\mu}_{z_r|s}$ and $\boldsymbol{C}_{z_r|s}$ is detailed. Since $\mathrm{E}\{\mathrm{Re}\{a\}\} = \mathrm{Re}\{\mathrm{E}\{a\}\}$, and $\mathrm{E}\{\mathrm{Im}\{a\}\} = \mathrm{Im}\{\mathrm{E}\{a\}\}$, first the expected value of the complex received signal is computed, which reads as

$$\mathrm{E}\left\{z_k[t]|s\right\} = \mathrm{E}\left\{\boldsymbol{h}_k\boldsymbol{x}[t]\ |s\right\}. \tag{5-14}$$

To simplify the notation $y_k(\boldsymbol{s}) = \boldsymbol{h}_k\ \boldsymbol{x}\left(\boldsymbol{s}\right)$ is introduced. The mean vector $\boldsymbol{\mu}_{z_r|s}$ is, then, given by

$$\boldsymbol{\mu}_{z_r|s} = \left[\mathrm{E}\left\{\mathrm{Re}\left\{y_k\left(\boldsymbol{s}\right)\right\}|s\right\} \quad \mathrm{E}\left\{\mathrm{Im}\left\{y_k\left(\boldsymbol{s}\right)\right\}|s\right\}\right]^T, \tag{5-15}$$

where

$$\mathrm{E}\left\{\mathrm{Re}\left\{y_k\left(\boldsymbol{s}\right)\right\}|s\right\} = \left(\frac{1}{\alpha_s}\right)^{K-1}\sum_{\boldsymbol{s}\in\mathcal{D}}\mathrm{Re}\left\{y_k\left(\boldsymbol{s}\right)\right\}, \tag{5-16}$$

$$\mathrm{E}\left\{\mathrm{Im}\left\{y_k\left(\boldsymbol{s}\right)\right\}|s\right\} = \left(\frac{1}{\alpha_s}\right)^{K-1}\sum_{\boldsymbol{s}\in\mathcal{D}}\mathrm{Im}\left\{y_k\left(\boldsymbol{s}\right)\right\} \tag{5-17}$$

and $\mathcal{D}$ is the set of all possible $\boldsymbol{s}[t]$ whose k-th entry is s. Moreover, the corresponding covariance matrix is given by

$$\boldsymbol{C}_{z_r|s} = \begin{bmatrix}\sigma_{r|s}^2 & \rho_{ri|s}\\ \rho_{ri|s} & \sigma_{i|s}^2\end{bmatrix}. \tag{5-18}$$

The entries of $\boldsymbol{C}_{z_r|s}$ read as

$$\sigma^2_{r|s} = \frac{\sigma^2_w}{2} + \mathrm{E}\left\{\mathrm{Re}\left\{y_k\left(\boldsymbol{s}\right)\right\}^2|s\right\} - \mathrm{E}\left\{\mathrm{Re}\left\{y_k\left(\boldsymbol{s}\right)\right\}|s\right\}^2, \tag{5-19}$$

$$\sigma^2_{i|s} = \frac{\sigma^2_w}{2} + \mathrm{E}\left\{\mathrm{Im}\left\{y_k\left(\boldsymbol{s}\right)\right\}^2|s\right\} - \mathrm{E}\left\{\mathrm{Im}\left\{y_k\left(\boldsymbol{s}\right)\right\}|s\right\}^2, \tag{5-20}$$

$$\rho_{ri|s} = \mathrm{E}\{\mathrm{Re}\{y_k\left(\boldsymbol{s}\right)\}\mathrm{Im}\{y_k\left(\boldsymbol{s}\right)|s\}\} - \mathrm{E}\{\mathrm{Re}\{y_k\left(\boldsymbol{s}\right)\}|s\}\mathrm{E}\{\mathrm{Im}\left\{y_k\left(\boldsymbol{s}\right)\right\}|s\}, \tag{5-21}$$

where

$$\mathrm{E}\left\{\mathrm{Re}\left\{y_k\left(\boldsymbol{s}\right)\right\}^2|s\right\} = \left(\frac{1}{\alpha_s}\right)^{K-1}\sum_{\boldsymbol{s}\in\mathcal{D}}\mathrm{Re}\left\{y_k\left(\boldsymbol{s}\right)\right\}^2, \tag{5-22}$$

$$\mathrm{E}\left\{\mathrm{Im}\left\{y_k\left(\boldsymbol{s}\right)\right\}^2|s\right\} = \left(\frac{1}{\alpha_s}\right)^{K-1}\sum_{\boldsymbol{s}\in\mathcal{D}}\mathrm{Im}\left\{y_k\left(\boldsymbol{s}\right)\right\}^2, \tag{5-23}$$

$$\mathrm{E}\left\{\mathrm{Re}\left\{y_k\left(\boldsymbol{s}\right)\right\}\mathrm{Im}\left\{y_k\left(\boldsymbol{s}\right)\right\}|s\right\} = \left(\frac{1}{\alpha_s}\right)^{K-1}\sum_{\boldsymbol{s}\in\mathcal{D}}\mathrm{Re}\left\{y_k\left(\boldsymbol{s}\right)\right\}\mathrm{Im}\left\{y_k\left(\boldsymbol{s}\right)\right\} \tag{5-24}$$

and $\mathrm{E}\left\{\mathrm{Re}\left\{y_k\left(\boldsymbol{s}\right)\right\}|s\right\}$ and $\mathrm{E}\left\{\mathrm{Im}\left\{y_k\left(\boldsymbol{s}\right)\right\}|s\right\}$ are defined in equations (5-16) and (5-17), respectively. Based on $\boldsymbol{C}_{z_r|s}$ and $\boldsymbol{\mu}_{z_r|s}$, $L_e(c_{k,i})$ is computed as

$$L_e(c_{k,i}) = \ln\left(\frac{\displaystyle\sum_{s\in S_{+1}}\frac{\mathrm{e}^{\Psi_s}}{\sqrt{\det\left(\boldsymbol{C}_{z_r|s}\right)}}P\left(s|r_{k,t,\upsilon}=+1\right)}{\displaystyle\sum_{s\in S_{-1}}\frac{\mathrm{e}^{\Psi_s}}{\sqrt{\det\left(\boldsymbol{C}_{z_r|s}\right)}}P\left(s|r_{k,t,\upsilon}=-1\right)}\right), \tag{5-25}$$

where

$$\Psi_s = -\frac{1}{2}\left[\left(\boldsymbol{z}_r[t]-\boldsymbol{\mu}_{z_r|s}\right)^T\boldsymbol{C}^{-1}_{z_r|s}\left(\boldsymbol{z}_r[t]-\boldsymbol{\mu}_{z_r|s}\right)\right] \tag{5-26}$$

and $P\left(s|r_{k,t,\upsilon}=g\right)$ for $g\in\{-1,+1\}$ can be computed with equation (5-6) considering $L_a(c_{k,i})$ for $i=1,\ldots,\frac{N_b}{R}$.

Note that, when calculating $L_e(c_{k,i})$ using (5-25), $\tilde{p}(z_k[t]|s)$ is evaluated only α_s times. This results in a significant decrease in computational complexity, when compared with the approach proposed in (5-12). However, for computing (5-25), the receiver requires access to $\boldsymbol{C}_{z_r|s}$ and $\boldsymbol{\mu}_{z_r|s}$ for all values of s. These parameters need to be provided by the BS which causes communication overhead. In this context, an alternative method that requires fewer number of parameters to be transmitted is desired.

5.1.4 Linear Model Based Discrete Precoding Aware Soft Detector

In this subsection, a method for computing $L_e(c_{k,i})$ with a reduced number of model parameters is devised. This proposed approach relies on the description of the received signal by a linear model.

5.1.4.1 Discrete Precoding Aware Linear Model

The Discrete Precoding Aware Linear Model (DPA-LM) is based on the assumption that the received signal can be expressed by

$$z_k[t] = h_k^{\text{eff}} s_k[t] + w_k[t] + \epsilon_k[t], \tag{5-27}$$

where $h_k^{\text{eff}} \in \mathcal{C}$ is a factor that expresses the precoder and channel effects on the transmit symbol of the k-th user and $\epsilon_k[t]$ is the error term that denotes the difference between $z_k[t]$ and $h_k^{\text{eff}} s_k[t] + w_k[t]$. To identify an appropriate h_k^{eff} the following MSE optimization problem is considered

$$\begin{aligned} h_k^{\text{eff}} &= \arg\min \lambda_{\epsilon_k}^2 = \arg\min \mathrm{E}\left\{|\epsilon_k[t]|^2\right\} \\ &= \arg\min_{\gamma \in \mathcal{C}} \mathrm{E}\left\{|\boldsymbol{h}_k\ \boldsymbol{x}[t] - \gamma\ s_k[t]|^2\right\}, \end{aligned} \tag{5-28}$$

where the optimal solution is given by

$$h_k^{\text{eff}} = \frac{1}{\alpha_s^K\ \sigma_s^2} \sum_{\boldsymbol{s} \in \mathcal{S}^K} s_k^*(\boldsymbol{s})\ y_k(\boldsymbol{s}), \tag{5-29}$$

$$\lambda_{\epsilon_k}^2 = \boldsymbol{h}_k\ \boldsymbol{\Lambda}_x\ \boldsymbol{h}_k^H - \left|h_k^{\text{eff}}\right|^2 \sigma_s^2, \tag{5-30}$$

where $\boldsymbol{\Lambda}_x = \left(\frac{1}{\alpha_s}\right)^K \sum_{\boldsymbol{s} \in \mathcal{S}^K} \boldsymbol{x}(\boldsymbol{s})\, \boldsymbol{x}(\boldsymbol{s})^H$ and $s_k(\boldsymbol{s})$ is the k-th element of $\boldsymbol{s}$. The derivation of the values for h_k^{eff} and $\lambda_{\epsilon_k}^2$ is given in the Appendix B.1.

5.1.4.2 DPA-LM Soft Detector

This subsection describes the proposed DPA-LM Soft Detector as a method for computing the extrinsic information based on the linear model previously presented.

The following strategy relies on the assumption that the error term $\epsilon_k[t]$ is a circular symmetric complex Gaussian random variable. The expected value

of the received signal is calculated as

$$\mathrm{E}\left\{z_k[t]|s\right\} = h_k^{\mathrm{eff}} s + \mathrm{E}\left\{\epsilon_k[t]|s\right\}, \tag{5-31}$$

and assuming $\mathrm{E}\left\{\epsilon_k[t]|s\right\} = 0 \ \forall \ s \in \mathcal{S}$ yields

$$\boldsymbol{\mu}_{z_r|s}^{\mathrm{eff}} = \left[\mathrm{Re}\left\{h_k^{\mathrm{eff}} s\right\} \quad \mathrm{Im}\left\{h_k^{\mathrm{eff}} s\right\}\right]^T. \tag{5-32}$$

Considering that

$$\boldsymbol{C}_{z_r}^{\mathrm{eff}} = \frac{\sigma_{\mathrm{eff}_k}^2}{2}\boldsymbol{I}, \tag{5-33}$$

with $\sigma_{\mathrm{eff}_k}^2 = \lambda_{\epsilon_k}^2 + \sigma_w^2$ being the effective noise variance. Then, the extrinsic information function from (5-4) simplifies to

$$L_e\left(c_{k,i}\right) = \ln\left(\frac{\sum\limits_{s\in S_{+1}} \mathrm{e}^{-\frac{\left|z_k[t]-h_k^{\mathrm{eff}}\ s\right|^2}{\sigma_{\mathrm{eff}_k}^2}} P\left(s|r_{k,t,\upsilon}=+1\right)}{\sum\limits_{s\in S_{-1}} \mathrm{e}^{-\frac{\left|z_k[t]-h_k^{\mathrm{eff}}\ s\right|^2}{\sigma_{\mathrm{eff}_k}^2}} P\left(s|r_{k,t,\upsilon}=-1\right)}\right), \tag{5-34}$$

where the values for $P\left(s|r_{k,t,\upsilon}=g\right)$, for $g \in \{-1,+1\}$ are calculated using equation (5-6).

The computation of $L_e(c_{k,i})$ according to (5-34) only requires knowledge about the parameters h_k^{eff} and $\sigma_{\mathrm{eff}_k}^2$, which are independent of the data symbol s. In comparison with the method from subsection 5.1.3, the number of parameters that need to be transmitted in advance to the information data is significantly reduced.

5.2 DPA-IDD Algorithm

Subsections 5.1.2, 5.1.3 and 5.1.4.2 expose different methods for computing $L_e(c_{k,i})$ when $L_a(c_{k,i})$ is known. Using these results, the DPA-IDD scheme is presented as a way of computing $L(c_{k,i})$ via making an iterative estimation of $L_a(c_{k,i})$ and, consequently, $L_e(c_{k,i})$. For description of the DPA-IDD scheme the vectors $\boldsymbol{L}, \boldsymbol{L}_e$ and $\boldsymbol{L}_a$ are defined as

$$\boldsymbol{L} = \begin{bmatrix} L(c_{k,1}) \\ \vdots \\ L(c_{k,\frac{N_b}{R}}) \end{bmatrix} \quad \boldsymbol{L}_e = \begin{bmatrix} L_e(c_{k,1}) \\ \vdots \\ L_e(c_{k,\frac{N_b}{R}}) \end{bmatrix} \quad \boldsymbol{L}_a = \begin{bmatrix} L_a(c_{k,1}) \\ \vdots \\ L_a(c_{k,\frac{N_b}{R}}) \end{bmatrix}.$$

The principle of the proposed receiver is based on equation (5-2). Based on $\boldsymbol{L}$ and $\boldsymbol{L}_e$, the a priori information is extracted via $\boldsymbol{L}_a = \boldsymbol{L} - \boldsymbol{L}_e$.

With this, for initialization, the detector calculates $\boldsymbol{L}_e$ assuming $\boldsymbol{L}_a = \boldsymbol{0}$ and forwards it to the decoder. The decoder outputs the LLR vector $\boldsymbol{L}$. Using $\boldsymbol{L}$ and $\boldsymbol{L}_e$, the a priori information is calculated and fed back into the detector which will, then, recompute $\boldsymbol{L}_e$ based on the updated $\boldsymbol{L}_a$. This process is done recursively until the maximum number of iterations is reached. An illustration of the receiving process is shown in Fig. 5.1.

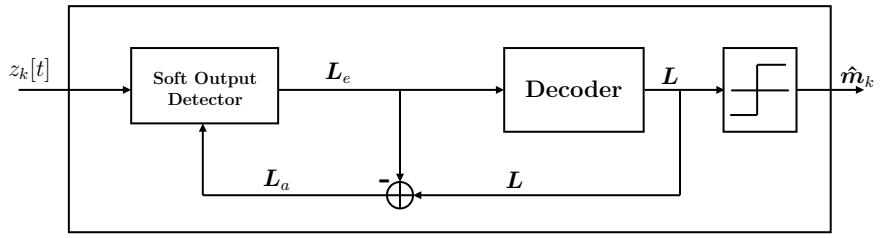

Figure 5.1: DPA-IDD Receiver Topology

The DPA-IDD technique does not require a specific method for computing $\boldsymbol{L}_e$. Hence, the approaches presented in subsections 5.1.2, 5.1.3 and 5.1.4.2 are compatible with the framework and can be used for calculating $\boldsymbol{L}_e$. The steps of the proposed approach are further detailed in Algorithm 4.

Note that, unlike other IDD approaches, e.g. [44], the proposed soft detectors compute $L_e(c_{k,i})$ instead of $L(c_{k,i})$. As a consequence, there is no need for subtracting the a priori information from the soft detector's output, as shown in Fig. 5.1.

Algorithm 4 Proposed DPA-IDD Algorithm

Select one method for the extrinsic information computation
Initialize $\boldsymbol{L}_a = \mathbf{0}$
for $j = 1 : N_{iter} - 1$ **do**
  for $i = 1 : \frac{N_b}{R}$ **do**
    Using $\boldsymbol{L}_a$, compute $L_e(c_{k,i})$ according to the method previously selected
  end for
  Organize $\boldsymbol{L}_e = \left[L_e(c_{k,1}), \ldots, L_e(c_{k,\frac{N_b}{R}})\right]^T$
  Forward $\boldsymbol{L}_e$ to the message-passing decoder
  Using the decoder's output $\boldsymbol{L}$, update $\boldsymbol{L}_a = \boldsymbol{L} - \boldsymbol{L}_e$
end for
for $i = 1 : \frac{N_b}{R}$ **do**
  Using $\boldsymbol{L}_a$, compute $L_e(c_{k,i})$ according to the method previously selected
end for
Organize $\boldsymbol{L}_e = \left[L_e(c_{k,1}), \ldots, L_e(c_{k,\frac{N_b}{R}})\right]^T$
Forward $\boldsymbol{L}_e$ to the message-passing decoder
Using the decoder's output compute $\hat{\boldsymbol{c}}_k = \text{sgn}(\boldsymbol{L})$
Finally $\hat{\boldsymbol{m}} = \left[c_{k,\frac{N_b(1-R)}{R}+1}, \ldots, c_{k,\frac{N_b}{R}}\right]$

6
Numerical Results

For the numerical evaluation, the BER is considered. It is assumed that the channel gains are modeled by independent Rayleigh fading [45], meaning $\beta_m = 1$ for $m = 1, \ldots, M$ and $g_{k,m} \sim \mathcal{CN}(0, \sigma_g^2)$ for $k = 1, \ldots, K$ and $m = 1, \ldots, M$ as done implicitly in [11] and [20] and explicitly in [18]. Moreover, the SNR is defined by $\text{SNR} = \frac{\|\boldsymbol{x}\|_2^2}{N_0}$, where the spectral noise power density N_0 is equivalent to the noise sample variance σ_w^2.

This section is divided into two different parts. In the first part, an uncoded transmission is evaluated where hard detection is used. In the second part, a coded transmission is considered and the proposed soft detection schemes are evaluated against the common method for AWGN channels described in subsection 5.1.1.

The analysis is made in conjunction with the IDD scheme presented in subsection 5.2. It relies on the utilization of the proposed MMSE branch-and-bound technique as the precoding method.

6.1
Uncoded Transmission

In this subsection the performance of the proposed precoding algorithms are evaluated against with following the state-of-the-art approaches:

1. The MSM-Precoder [20] considering phase quantization which implies solving an LP;
2. The ZF precoder with constant envelope [11], where the entries of the precoding vector are subsequently phase quantized;
3. The phase quantized CIO precoder implemented via CVX [18], which corresponds to solving a second order cone program.

In this context, the subsection is divided into two parts. In the first part, the precoding strategies are evaluated in terms of BER using phase quantizers as hard detectors. In the second part, the complexity of the proposed methods is analyzed and compared against the mentioned algorithms.

6.1.1 Hard Detection using Phase Quantizers at the Receiver

In this subsection, the performance of the proposed algorithms is evaluated with phase quantizers as hard detectors. The analyzed case considers, both, the data and the transmit vector symbols are 8-PSK, which means $\alpha_s = 8$ and $\alpha_x = 8$. Two different scenarios are considered. First the BER performance of the proposed MMSE and MMDDT branch-and-bound algorithms is compared. The considered system has $K = 2$ users and $M = 4$ BS antennas. The BER performances are illustrated in Fig. 6.1.

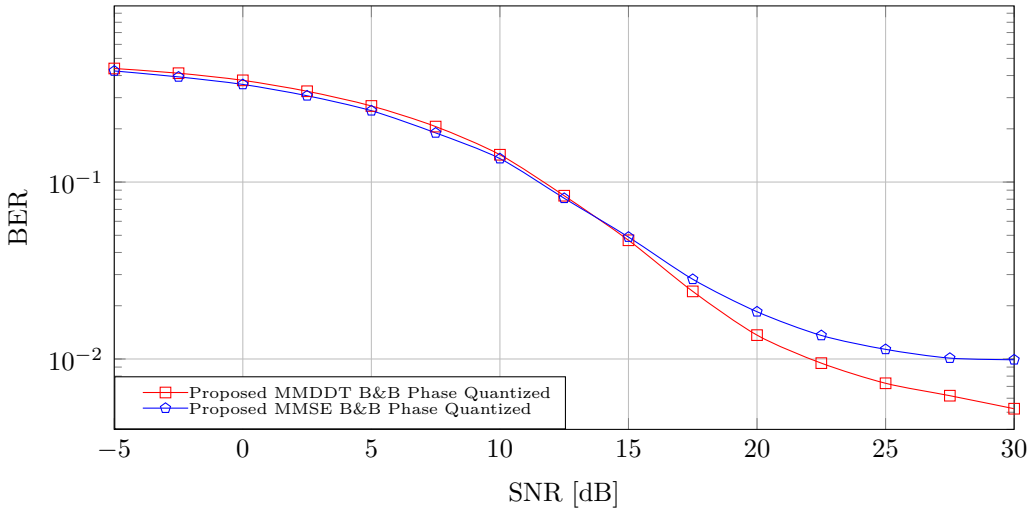

Figure 6.1: Uncoded BER versus SNR for $K = 2$, $M = 4$, $\alpha_s = 8$ and $\alpha_x = 8$

In the second scenario, the proposed methods are compared with state-of-the-art approaches for a system with $K = 3$ users and $M = 12$ BS antennas. The BER performances are illustrated in Fig. 6.2.

Fig. 6.1 confirms the superiority of the MMSE criterion against MMDDT for low SNR. On the other hand, Fig. 6.1 shows that the MMDDT criterion is favorable for the high SNR regime where it yields a lower uncoded BER.

The results shown in Fig. 6.2 illustrate a significant gain in BER performance when using the optimal branch-and-bound methods in contrast to suboptimal low resolution approaches. Moreover, the proposed branch-and-bound precoders shows only a 2 dB loss in comparison with the full resolution MMSE linear precoding strategy presented in [13].

Fig. 6.2 also confirms the suitability of the MMSE criterion for low and medium SNR once the proposed optimal approach outperforms all other low resolution schemes for that SNR regime, in terms of BER. Besides that, the results indicate that the proposed suboptimal approach termed MMSE

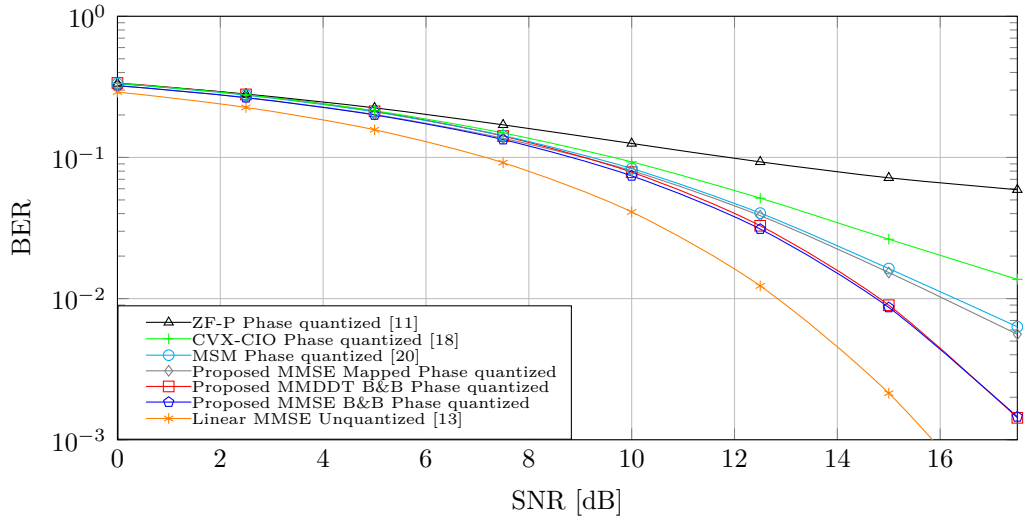

Figure 6.2: Uncoded BER versus SNR, $K = 3$, $M = 12$, $\alpha_s = 8$ and $\alpha_x = 8$

Mapped surprisingly outperforms other suboptimal state-of-the-art algorithms in terms of BER performance.

6.1.2 Complexity Analysis

In this subsection the computational complexity of the proposed algorithms is evaluated and compared against the state-of-the-art approaches. For the analysis considered the system has $K = 3$ users and uses QPSK modulation for both the data and transmit vector, meaning $\alpha_x = \alpha_s = 4$.

The computational complexity of each algorithm is summarized in Table 6.1, where B denotes the number of evaluated bounds in the corresponding branch-and-bound algorithm.

Table 6.1: Computational Complexity of the Algorithms

Algorithm	Complexity
MSM-Precoder [20]	$\mathcal{O}((2M+1)^{3.5})$
ZF-P [11]	$\mathcal{O}(K^2M)$
CVX-CIO [18]	$\mathcal{O}((2M+1)^{3.5})$
RedMinBER [22]	$\mathcal{O}(N_p\tilde{M}K)$
Proposed MMDDT branch-and-bound	$\mathcal{O}(B\,(2M+1)^{3.5})$
Proposed MMSE branch-and-bound	$\mathcal{O}(B\,(2M+1)^{3.5})$
Proposed MMSE Mapped	$\mathcal{O}((2M+1)^{3.5})$

The results shown in Table 6.1 were obtained considering that by using the interior points method (IPM) the optimization problems of the state-of-the-art algorithms can be solved with complexity in the order of $\mathcal{O}(n^{3.5})$, with $n \leq (2M+1)$, cf. [42]. Furthermore, for the RedMinBER precoder it

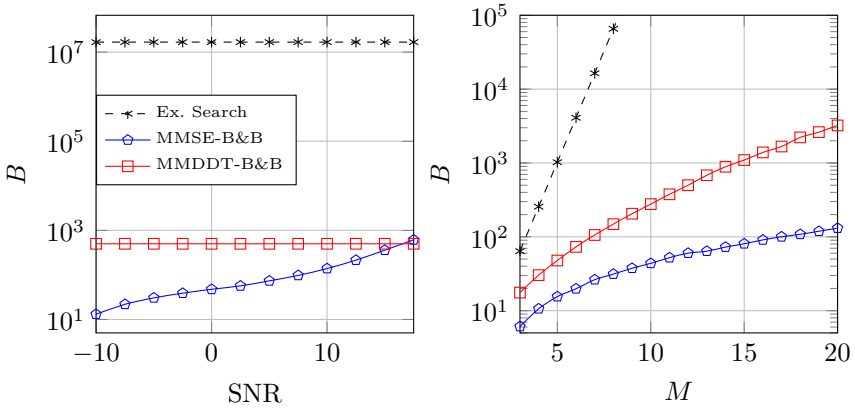

Figure 6.3: Average # of evaluated bounds × SNR, $K = 3$, $M = 12$, $\alpha_s = \alpha_x = 4$ (left). Average # of evaluated bounds × Number of transmit antennas, $K = 3$, SNR = 3dB, $\alpha_s = \alpha_x = 4$ (right).

is considered that the gradient descent method uses N_p iterations with each one having complexity $\tilde{M}K$, where typically $\tilde{M} \ll M$.

Fig. 6.3 shows the number of evaluated bounds of the proposed branch-and-bound methods compared exhaustive search for the considered system. In Fig. 6.3 it can be seen that the number of evaluated bounds of the proposed MMSE branch-and-bound method is significantly smaller than the one from MMDDT for low SNR, which underlines the superiority of the MMSE criterion for low SNR. However, for the high-SNR regime the complexity of the proposed MMDDT branch-and-bound method is smaller than the one from the proposed MMSE approach, which underlines the superiority of the MMDDT criterion for high SNR.

Furthermore, based on Fig. 6.3, the average number of subproblems solved using the proposed methods is always significantly smaller than the total number of candidates to be evaluated in the exhaustive search. Taking into account that each candidate evaluation in the exhaustive search corresponds to a complexity of $\mathcal{O}(K^2M)$ justifies the utilization of the proposed method when the optimal precoding vector is desired.

6.2 Coded Transmission

In this subsection, the proposed soft detection schemes are evaluated considering as the proposed MMSE branch-and-bound approach the precoding technique.

The shown results were computed using an LDPC block code with a block size of $\frac{N_b}{R} = 486$ bits and code rate $R = 1/2$. The LLRs are processed by sum-product algorithm (SPA) decoders [46]. The examined system has $K = 2$ users and $B = 6$ BS antennas where the data symbols are considered as 8-

PSK and the precoded symbols are considered as QPSK, meaning $\alpha_s = 8$ and $\alpha_x = 4$. The examined approaches are as follows

1. Uncoded transmission
2. Coded transmission using the DPA soft detector (5-12) for the computation of $\boldsymbol{L}_e$;
3. Coded transmission using the GDPA soft detector (5-25) for the computation of $\boldsymbol{L}_e$;
4. Coded transmission using DPA-LM soft detector (5-34) for the computation of $\boldsymbol{L}_e$;
5. Coded transmission using AWGN method (5-9) for the computation of $\boldsymbol{L}_e$.

Note that the uncoded system provides higher data rate than the coded versions.

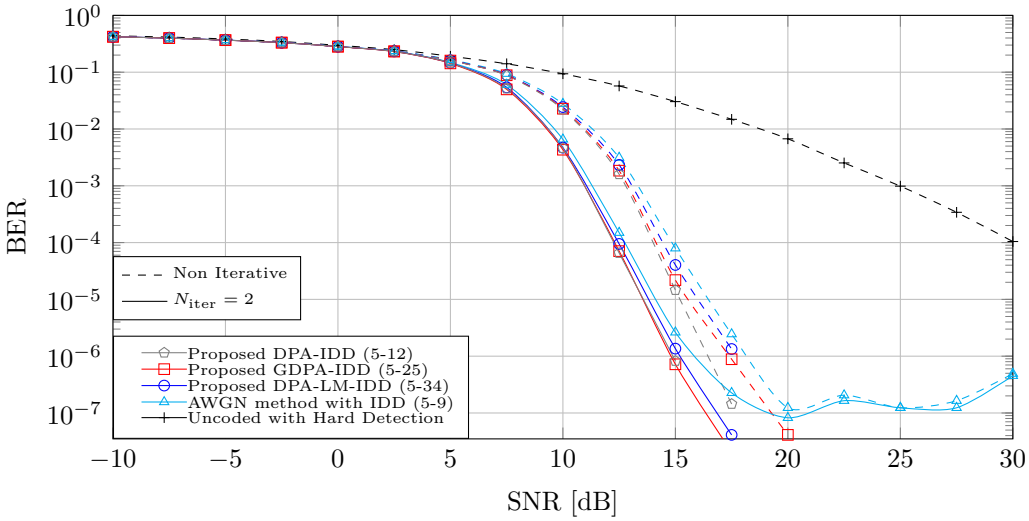

Figure 6.4: Coded BER versus SNR, $K = 2$, $M = 6$, $\alpha_s = 8$, $\alpha_x = 4$

As can be seen in Fig. 6.4, all proposed methods provide similar performance for low-SNR. As expected, for the high-SNR regime the proposed DPA-IDD method, that relies on the true PDF of the received signal, yields a lower BER as compared with the proposed suboptimal methods. Furthermore, considering the marginal performance loss referring to the proposed DPA-IDD method, shown in Fig. 6.4, reasonable complexity performance trade-offs can be achieved via using the proposed suboptimal methods GDPA-IDD and DPA-LM-IDD.

The BER performance associated with the system that uses the common AWGN soft detector is similar to the proposed methods for low-SNR. However,

in the medium and high-SNR regime, the distortion brought by the discrete precoding becomes relevant, and, since this is not considered in the common AWGN receive processing it results in an error floor in the BER, as shown in Fig. 6.4.

Finally, Fig. 6.4 shows an improvement in performance when using the iterative method. With a relatively small number of iterations there is a gain of approximately 1.5 dB when compared with the non iterative approach.

7
Conclusions

In this study, two novel optimal precoding branch-and-bound algorithms and its corresponding suboptimal methods constrained to CE signals and phase quantization were devised. The first maximizes the MDDT at the receivers, while the second minimizes the MSE between the users' data symbols and the receive signal.

Moreover, three different soft detection methods and an iterative detection and decoding scheme that allow the utilization of channel coding in conjunction with low-resolution precoding were proposed. Besides an exact approach for computing the extrinsic information, two approximations with reduced computational complexity were devised.

The proposed branch-and-bound precoding algorithms are superior to the existing state-of-the-art methods in terms of uncoded BER. Moreover numerical results indicate that the proposed suboptimal MMSE design outperforms marginally the other examined suboptimal methods in terms of uncoded BER. Numerical results also show that the proposed approaches also have significantly lower complexity than exhaustive search. Finally, results based on an LDPC block code indicated that the proposed receive processing schemes yield a lower BER when compared to the conventional design.

8
Future Work

First, an interesting topic of study is the investigation of the sum-rate and energy efficiency of the proposed system. Comparison with other state-of-the-art methods in these regards is also desired.

For this study the data and precoding modulation were considered as PSK. An open issue for future work is the extension of the proposed precoding and detection approaches for different kinds of modulation e.g. QAM. Moreover, an extension of the proposed DPA-IDD approach for multiple receive antenna systems is desired.

Another open issue is the study of the effect of imperfect CSI on the performance of the proposed precoding and detection approaches. In this context, the development of receivers and precoders that consider imperfect CSI in its formulation is desired.

For reducing the computational complexity of the branch-and-bound based precoders it would be interesting to study different types of search methods, such as depth first search and best first search. In the context of reducing the computational cost of the branch-and-bound approaches it would be also of great interest to study different node enumeration techniques.

Finally, the investigation and development of true constant envelope precoding techniques is desired to enable the utilization of PAs with even higher efficiency.

Bibliography

[1] KATALINIC, A.; NAGY, R. ; ZENTNER, R.. **Benefits of mimo systems in practice: Increased capacity, reliability and spectrum efficiency**. In: PROCEEDINGS ELMAR 2006, p. 263–266, 2006.

[2] KHAN, L. U.; YAQOOB, I.; IMRAN, M.; HAN, Z. ; HONG, C. S.. **6G Wireless Systems: A Vision, Architectural Elements, and Future Directions**. IEEE Access, 8:147029–147044, 2020.

[3] RUSEK, F.; PERSSON, D.; LAU, B. K.; LARSSON, E. G.; MARZETTA, T. L.; EDFORS, O. ; TUFVESSON, F.. **Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays**. IEEE Signal Processing Magazine, 30(1):40–60, 2013.

[4] GIORDANI, M.; POLESE, M.; MEZZAVILLA, M.; RANGAN, S. ; ZORZI, M.. **Toward 6G Networks: Use Cases and Technologies**. IEEE Communications Magazine, 58(3):55–61, 2020.

[5] ELMEADAWY, S.; SHUBAIR, R. M.. **6G Wireless Communications: Future Technologies and Research Challenges**. In: 2019 INTERNATIONAL CONFERENCE ON ELECTRICAL AND COMPUTING TECHNOLOGIES AND APPLICATIONS (ICECTA), p. 1–5, 2019.

[6] YANG, P.; XIAO, Y.; XIAO, M. ; LI, S.. **6G Wireless Communications: Vision and Potential Techniques**. IEEE Network, 33(4):70–75, 2019.

[7] CRIPPS, S.. **RF Power Amplifiers for Wireless communications**. 1999.

[8] WANG, F.; LI, T.; HU, S. ; WANG, H.. **A Super-Resolution Mixed-Signal Doherty Power Amplifier for Simultaneous Linearity and Efficiency Enhancement**. IEEE Journal of Solid-State Circuits, 54(12):3421–3436, 2019.

[9] WALDEN, R.. **Analog-to-digital converter survey and analysis**. IEEE J. Sel. Areas Commun., 17(4):539 –550, Apr. 1999.

[10] JACOBSSON, S.; DURISI, G.; COLDREY, M.; GOLDSTEIN, T. ; STUDER, C.. **Quantized Precoding for Massive MU-MIMO**. IEEE Transactions on Communications, 65(11):4670–4684, 2017.

[11] MOHAMMED, S. K.; LARSSON, E. G.. **Per-antenna constant envelope precoding for large multi-user MIMO systems**. IEEE Trans. Commun., 61(3):1059–1071, March 2013.

[12] JACOBSSON, S.; DURISI, G.; COLDREY, M.; GOLDSTEIN, T. ; STUDER, C.. **Quantized Precoding for Massive MU-MIMO**. IEEE Trans. Commun., 65(11):4670–4684, 2017.

[13] JOHAM, M.; UTSCHICK, W. ; NOSSEK, J. A.. **Linear transmit processing in MIMO communications systems**. IEEE Trans. Signal Process., 53(8):2700–2712, Aug 2005.

[14] MEZGHANI, A.; GHIAT, R. ; NOSSEK, J. A.. **Transmit processing with low resolution D/A-converters**. In: 2009 16TH IEEE INTERNATIONAL CONFERENCE ON ELECTRONICS, CIRCUITS AND SYSTEMS - (ICECS 2009), p. 683–686, Dec 2009.

[15] SAXENA, A. K.; FIJALKOW, I. ; SWINDLEHURST, A. L.. **On one-bit quantized ZF precoding for the multiuser massive MIMO downlink**. In: 2016 IEEE SENSOR ARRAY AND MULTICHANNEL SIGNAL PROCESSING WORKSHOP (SAM), p. 1–5, July 2016.

[16] SAXENA, A. K.; FIJALKOW, I. ; SWINDLEHURST, A. L.. **Analysis of One-Bit Quantized Precoding for the Multiuser Massive MIMO Downlink**. IEEE Trans. Signal Process., 65(17):4624–4634, 2017.

[17] JACOBSSON, S.; DURISI, G.; COLDREY, M.; GOLDSTEIN, T. ; STUDER, C.. **Nonlinear 1-bit precoding for massive MU-MIMO with higher-order modulation**. In: 2016 50TH ASILOMAR CONFERENCE ON SIGNALS, SYSTEMS AND COMPUTERS, p. 763–767, 2016.

[18] AMADORI, P. V.; MASOUROS, C.. **Constant envelope precoding by interference exploitation in phase shift keying-modulated multiuser transmission**. IEEE Trans. Commun., 16(1):538–550, Jan 2017.

[19] NOLL, A.; JEDDA, H. ; NOSSEK, J.. **PSK Precoding in Multi-User MISO Systems**. In: WSA 2017; 21TH INTERNATIONAL ITG WORKSHOP ON SMART ANTENNAS, p. 1–7, Berlin, Germany, 2017.

[20] JEDDA, H.; MEZGHANI, A.; SWINDLEHURST, A. L. ; NOSSEK, J. A.. **Quantized constant envelope precoding with PSK and QAM signaling**. IEEE Trans. Wireless Commun., 17(12):8022–8034, Dec 2018.

[21] CHU, L.; WEN, F.; LI, L. ; QIU, R.. **Efficient Nonlinear Precoding for Massive MIMO Downlink Systems With 1-Bit DACs**. IEEE Trans. Wireless Commun., 18(9):4213–4224, 2019.

[22] LEE SWINDLEHURST, A.; JEDDA, H. ; FIJALKOW, I.. **Reduced Dimension Minimum BER PSK Precoding for Constrained Transmit Signals in Massive MIMO**. In: 2018 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), p. 3584–3588, 2018.

[23] FESL, B.; JEDDA, H. ; NOSSEK, J. A.. **Discrete One-Bit Precoding for Massive MIMO**. In: WSA 2019; 23RD INTERNATIONAL ITG WORKSHOP ON SMART ANTENNAS, p. 1–6, April 2019.

[24] NEDELCU, A.; STEINER, F.; STAUDACHER, M.; KRAMER, G.; ZIRWAS, W.; GANESAN, R. S.; BARACCA, P. ; WESEMANN, S.. **Quantized precoding for multi-antenna downlink channels with MAGIQ**. In: PROC. OF THE 22ND INT. ITG WORKSHOP ON SMART ANTENNAS, Bochum, Germany, March 2018.

[25] LANDAU, L. T. N.; DE LAMARE, R. C.. **Branch-and-bound precoding for multiuser MIMO systems with 1-bit quantization**. IEEE Wireless Commun. Lett., 6(6):770–773, Dec 2017.

[26] JACOBSSON, S.; XU, W.; DURISI, G. ; STUDER, C.. **MSE-optimal 1-bit precoding for multiuser MIMO via branch and bound**. In: PROC. IEEE INT. CONF. ACOUST., SPEECH, SIGNAL PROCESS., p. 3589–3593, Calgary, Alberta, Canada, April 2018.

[27] MEZGHANI, A.; KHOUFI, M. ; NOSSEK, J. A.. **Maximum likelihood detection for quantized MIMO systems**. In: 2008 INTERNATIONAL ITG WORKSHOP ON SMART ANTENNAS, p. 278–284, Vienna, Austria, 2008.

[28] RISI, C.; PERSSON, D. ; LARSSON, E. G.. **Massive MIMO with 1-bit ADC**. CoRR, abs/1404.7736, 2014.

[29] JEON, Y.; LEE, N.; HONG, S. ; HEATH, R. W.. **One-Bit Sphere Decoding for Uplink Massive MIMO Systems With One-Bit**

ADCs. IEEE Transactions on Wireless Communications, 17(7):4509–4521, 2018.

[30] SHAO, Z.; DE LAMARE, R. C. ; LANDAU, L. T. N.. **Iterative Detection and Decoding for Large-Scale Multiple-Antenna Systems With 1-Bit ADCs**. IEEE Wireless Communications Letters, 7(3):476–479, 2018.

[31] ZHANG, Z.; CAI, X.; LI, C.; ZHONG, C. ; DAI, H.. **One-Bit Quantized Massive MIMO Detection Based on Variational Approximate Message Passing**. IEEE Trans. Signal Process., 66(9):2358–2373, 2018.

[32] CHOI, J.; MO, J. ; HEATH, R. W.. **Near Maximum-Likelihood Detector and Channel Estimator for Uplink Multiuser Massive MIMO Systems With One-Bit ADCs**. IEEE Trans. Commun., 64(5):2005–2018, 2016.

[33] WANG, S.; LI, Y. ; WANG, J.. **Multiuser Detection in Massive Spatial Modulation MIMO With Low-Resolution ADCs**. IEEE Trans. Wireless Commun., 14(4):2156–2168, 2015.

[34] LOPES, E. S. P.; LANDAU, L. T. N.. **Optimal Precoding for Multiuser MIMO Systems With Phase Quantization and PSK Modulation via Branch-and-Bound**. IEEE Wireless Communications Letters, 2020.

[35] LOPES, E. S. P.; LANDAU, L. T. N.. **Optimal and Suboptimal MMSE Precoding for Multiuser MIMO Systems Using Constant Envelope Signals with Phase Quantization at the Transmitter and PSK**. In: WSA 2020; 24TH INTERNATIONAL ITG WORKSHOP ON SMART ANTENNAS, p. 1–6, 2020.

[36] LOPES, E. S. P.; LANDAU, L. T. N.. **Discrete MMSE Precoding for Multiuser MIMO Systems with PSK Modulation**, 2021.

[37] LOPES, E. S. P.; LANDAU, L. T. N.. **Iterative Detection and Decoding for Multiuser MIMO Systems with Low Resolution Precoding and PSK Modulation**. In: 2021 IEEE STATISTICAL SIGNAL PROCESSING WORKSHOP (SSP), p. 356–360, 2021.

[38] A. H. LAND AND A. G. DOIG. **An Automatic Method of Solving Discrete Programming Problems**. Econometrica, 28(3):497–520, 1960.

[39] LUO, J.; PATTIPATI, K.; WILLETT, P. ; BRUNEL, L.. **Branch-and-bound-based fast optimal algorithm for multiuser detection in**

synchronous CDMA. In: PROC. IEEE INT. CONF. COMMUN. (ICC), p. 3336–3340 vol.5, 2003.

[40] ISRAEL, J.; FISCHER, A.; MARTINOVIC, J.; JORSWIECK, E. A. ; MESYAGUTOV, M.. **Discrete Receive Beamforming**. IEEE Signal Process. Lett., 22(7):958–962, 2015.

[41] ISRAEL, J.; FISCHER, A. ; MARTINOVIC, J.. **A Branch-and-Bound Algorithm for Discrete Receive Beamforming with Improved Bounds**. In: 2015 IEEE INTERNATIONAL CONFERENCE ON UBIQUITOUS WIRELESS BROADBAND (ICUWB), p. 1–5, 2015.

[42] BOYD, S.; VANDENBERGHE, L.. **Convex Optimization**. Cambridge University Press, New York, NY, USA, 2004.

[43] JEDDA, H.; NOSSEK, J. A. ; MEZGHANI, A.. **Minimum BER precoding in 1-bit massive MIMO systems**. In: PROC. OF IEEE SENSOR ARRAY AND MULTICHANNEL SIGNAL PROCESSING WORKSHOP (SAM), Rio de Janeiro, Brazil, July 2016.

[44] TEN BRINK, S.; SPEIDEL, J. ; RAN-HONG YAN. **Iterative demapping and decoding for multilevel modulation**. In: IEEE GLOBECOM 1998 (CAT. NO. 98CH36250), volumen 1, p. 579–584 vol.1, 1998.

[45] YANG, H.; MARZETTA, T. L.. **Performance of conjugate and zero-forcing beamforming in large-scale antenna systems**. IEEE Journal on Selected Areas in Communications, 31(2):172–179, 2013.

[46] KSCHISCHANG, F. R.; FREY, B. J. ; LOELIGER, H. A.. **Factor graphs and the sum-product algorithm**. IEEE Trans. Inf. Theory, 47(2):498–519, 2001.

A
Convexity Analysis

A.1
The Conventional MMSE Cost Function With the Scaling Factor

The corresponding real valued function of the equivalent MMSE cost function including the scaling factor reads as

$$J(\boldsymbol{x}_{\mathrm{r}}, f) = f^2 \boldsymbol{x}_{\mathrm{r}}^T \boldsymbol{H}_{\mathrm{r}}^T \boldsymbol{H}_{\mathrm{r}} \boldsymbol{x}_{\mathrm{r}} - 2f \boldsymbol{x}_{\mathrm{r}}^T \boldsymbol{H}_{\mathrm{r}}^T \boldsymbol{s}_{\mathrm{r}} + f^2 \mathrm{E}\{\boldsymbol{w}_{\mathrm{r}}^T \boldsymbol{w}_{\mathrm{r}}\}. \tag{A-1}$$

The Hessian is constructed based on the partial derivatives given by

$$\begin{aligned} \boldsymbol{\Gamma} &= \frac{\partial^2 J(\boldsymbol{x}_{\mathrm{r}}, f)}{\partial \boldsymbol{x}_{\mathrm{r}} \partial \boldsymbol{x}_{\mathrm{r}}^T} = 2f^2 \boldsymbol{H}_{\mathrm{r}}^T \boldsymbol{H}_{\mathrm{r}}, \\ \epsilon &= \frac{\partial^2 J(\boldsymbol{x}_{\mathrm{r}}, f)}{\partial f^2} = 2\|\boldsymbol{H}_{\mathrm{r}} \boldsymbol{x}_{\mathrm{r}}\|_2^2 + 2\mathrm{E}\{\boldsymbol{w}_{\mathrm{r}}^T \boldsymbol{w}_{\mathrm{r}}\} \geq 0, \\ \boldsymbol{\eta} &= \frac{\partial^2 J(\boldsymbol{x}_{\mathrm{r}}, f)}{\partial \boldsymbol{x}_{\mathrm{r}} \partial f} = 4f \boldsymbol{H}_{\mathrm{r}}^T \boldsymbol{H}_{\mathrm{r}} \boldsymbol{x}_{\mathrm{r}} - 2\boldsymbol{H}_{\mathrm{r}}^T \boldsymbol{s}_{\mathrm{r}}. \end{aligned} \tag{A-2}$$

Positive semi-definiteness of the Hessian is established when the following inequality holds

$$\boldsymbol{v}^T \boldsymbol{\Gamma} \boldsymbol{v} + \nu^2 \epsilon + 2\nu \boldsymbol{\eta}^T \boldsymbol{v} \geq 0 \quad \text{for all } \nu, \boldsymbol{v}. \tag{A-3}$$

Assuming that $\boldsymbol{\Gamma} \succ \mathbf{0}$, the minimum value of the LHS of (A-3) is given by $\nu^2 \epsilon - \nu^2 \boldsymbol{\eta}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{\eta}$. For positive semi-definiteness of the Hessian, this minimum must be greater or equal to zero which yields the condition

$$\epsilon \geq \boldsymbol{\eta}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{\eta}. \tag{A-4}$$

Inserting the partial derivatives in (A-4) gives

$$2\|\boldsymbol{H}_{\mathrm{r}}\boldsymbol{x}_{\mathrm{r}}\|_2^2 + 2\mathrm{E}\{\boldsymbol{w}_{\mathrm{r}}^T\boldsymbol{w}_{\mathrm{r}}\} \geq \frac{1}{2f^2}\boldsymbol{\eta}^T(\boldsymbol{H}_{\mathrm{r}}^T\boldsymbol{H}_{\mathrm{r}})^{-1}\boldsymbol{\eta}, \tag{A-5}$$

$$2\|\boldsymbol{H}_{\mathrm{r}}\boldsymbol{x}_{\mathrm{r}}\|_2^2 + 2\mathrm{E}\{\boldsymbol{w}_{\mathrm{r}}^T\boldsymbol{w}_{\mathrm{r}}\} \geq 8\boldsymbol{x}_{\mathrm{r}}^T\boldsymbol{H}_{\mathrm{r}}^T\boldsymbol{H}_{\mathrm{r}}\boldsymbol{x}_{\mathrm{r}} - \frac{8}{f}\boldsymbol{x}_{\mathrm{r}}^T\boldsymbol{H}_{\mathrm{r}}^T\boldsymbol{s}_{\mathrm{r}} + 2\boldsymbol{s}_{\mathrm{r}}^T\boldsymbol{H}_{\mathrm{r}}(\boldsymbol{H}_{\mathrm{r}}^T\boldsymbol{H}_{\mathrm{r}})^{-1}\boldsymbol{H}_{\mathrm{r}}^T\boldsymbol{s}_{\mathrm{r}}, \tag{A-6}$$

which can be rearranged as

$$\frac{8}{f}\boldsymbol{x}_{\mathrm{r}}^T\boldsymbol{H}_{\mathrm{r}}^T\boldsymbol{s}_{\mathrm{r}} \geq 6\boldsymbol{x}_{\mathrm{r}}^T\boldsymbol{H}_{\mathrm{r}}^T\boldsymbol{H}_{\mathrm{r}}\boldsymbol{x}_{\mathrm{r}} + 2\boldsymbol{s}_{\mathrm{r}}^T\boldsymbol{H}_{\mathrm{r}}(\boldsymbol{H}_{\mathrm{r}}^T\boldsymbol{H}_{\mathrm{r}})^{-1}\boldsymbol{H}_{\mathrm{r}}^T\boldsymbol{s}_{\mathrm{r}} - 2\mathrm{E}\{\boldsymbol{w}_{\mathrm{r}}^T\boldsymbol{w}_{\mathrm{r}}\}.$$

The MMSE cost function is, then, in general not jointly convex in $\boldsymbol{x}_{\mathrm{r}}$ and f.

A.2 The Partial MMSE Cost Function

The MMSE cost function for the problem formulation with the f' scaled polyhedron and partially fixed precoding vector is given by

$$J\left(\boldsymbol{x}'_{\mathrm{r},f}, f'\right) = \left\|\boldsymbol{H}'_{\mathrm{r}}\boldsymbol{x}'_{\mathrm{r},f} - \boldsymbol{s}_{\mathrm{r}} + f'\boldsymbol{H}_{\mathrm{r,\,fixed}}\boldsymbol{x}_{\mathrm{r,\,fixed}}\right\|_2^2 + f'^2\mathrm{E}\{\boldsymbol{w}_{\mathrm{r}}^T\boldsymbol{w}_{\mathrm{r}}\}. \tag{A-7}$$

As in the previous subsection the Hessian can be constructed with the partial derivatives which are now given by

$$\begin{aligned}
\boldsymbol{\Gamma} &= \frac{\partial^2 J(\boldsymbol{x}'_{\mathrm{r},f}, f')}{\partial\boldsymbol{x}'_{\mathrm{r},f}\partial\boldsymbol{x}'^T_{\mathrm{r},f}} = 2\boldsymbol{H}'^T_{\mathrm{r}}\boldsymbol{H}'_{\mathrm{r}}, \\
\epsilon &= \frac{\partial^2 J(\boldsymbol{x}'_{\mathrm{r},f}, f')}{\partial f'^{\,2}} = 2\|\boldsymbol{H}_{\mathrm{r,\,fixed}}\boldsymbol{x}_{\mathrm{r,\,fixed}}\|_2^2 + 2\mathrm{E}\{\boldsymbol{w}_{\mathrm{r}}^T\boldsymbol{w}_{\mathrm{r}}\} \geq 0, \\
\boldsymbol{\eta} &= \frac{\partial^2 J(\boldsymbol{x}'_{\mathrm{r},f}, f')}{\partial\boldsymbol{x}'_{\mathrm{r},f}\partial f'} = 2\boldsymbol{H}'^T_{\mathrm{r}}\boldsymbol{H}_{\mathrm{r,\,fixed}}\boldsymbol{x}_{\mathrm{r,\,fixed}}.
\end{aligned} \tag{A-8}$$

Analogous to the previous subsection $\boldsymbol{\Gamma} \succ \mathbf{0}$ is assumed and then (A-4) is a sufficient condition for convexity. In this case, (A-4), after including the partial derivatives (A-8), can be rearranged to

$$\mathrm{E}\{\boldsymbol{w}_{\mathrm{r}}^T\boldsymbol{w}_{\mathrm{r}}\} \geq \boldsymbol{x}_{\mathrm{r,\,fixed}}^T\boldsymbol{H}_{\mathrm{r,\,fixed}}^T(\boldsymbol{H}'_{\mathrm{r}}(\boldsymbol{H}'^T_{\mathrm{r}}\boldsymbol{H}'_{\mathrm{r}})^{-1}\boldsymbol{H}'^T_{\mathrm{r}} - \boldsymbol{I})\boldsymbol{H}_{\mathrm{r,\,fixed}}\boldsymbol{x}_{\mathrm{r,\,fixed}}. \tag{A-9}$$

Convexity is established by showing that the RHS of (A-9) is always smaller or equal to zero. This can be shown by considering

$$\boldsymbol{v}^T(\boldsymbol{H}'_{\mathrm{r}}(\boldsymbol{H}'^T_{\mathrm{r}}\boldsymbol{H}'_{\mathrm{r}})^{-1}\boldsymbol{H}'^T_{\mathrm{r}} - \boldsymbol{I})\boldsymbol{v} \leq 0, \tag{A-10}$$

which holds for all $\boldsymbol{v}$, since $\boldsymbol{H}'_{\mathrm{r}}(\boldsymbol{H}'^{T}_{\mathrm{r}}\boldsymbol{H}'_{\mathrm{r}})^{-1}\boldsymbol{H}'^{T}_{\mathrm{r}}$ is a projection matrix where the eigenvalues can only be one or zero.

B
Linear Model Derivations

B.1
Derivation of h_k^{eff}

In the following, the expression for h_k^{eff} is derived. To minimize $\lambda_{\epsilon_k}^2$ the derivative of (5-28) with respect to γ^* is taken and equated to 0, which is expressed as

$$\frac{\partial \lambda_{\epsilon_k}^2}{\partial \gamma^*} = \gamma \, \mathrm{E}\left\{s_k(\boldsymbol{s}) \; s_k(\boldsymbol{s})^*\right\} - \mathrm{E}\left\{\boldsymbol{h}_k \; \boldsymbol{x}(\boldsymbol{s}) \; s_k(\boldsymbol{s})^*\right\} = 0. \tag{B-1}$$

With this, the effective channel coefficient reads as

$$h_k^{\text{eff}} = \gamma = \frac{\boldsymbol{h}_k \; \mathrm{E}\left\{s_k(\boldsymbol{s})^* \; \boldsymbol{x}(\boldsymbol{s})\right\}}{\sigma_s^2}. \tag{B-2}$$

Finally, h_k^{eff} is used for the computation of the mean squared error $\lambda_{\epsilon_k}^2 = \mathrm{E}\left\{|\epsilon_k[t]|^2\right\}$ as follows

$$\begin{aligned} \lambda_{\epsilon_k}^2 &= \boldsymbol{h}_k \mathrm{E}\left\{\boldsymbol{x}(\boldsymbol{s}) \; \boldsymbol{x}(\boldsymbol{s})^H\right\} \boldsymbol{h}_k^H - \; \frac{1}{\sigma_s^2} \left|\mathrm{E}\left\{\boldsymbol{h}_k \; \boldsymbol{x}(\boldsymbol{s}) \; k(\boldsymbol{s})^*\right\}\right|^2 \\ &= \boldsymbol{h}_k \; \boldsymbol{\Lambda}_x \; \boldsymbol{h}_k^H - \left|h_k^{\text{eff}}\right|^2 \sigma_s^2. \end{aligned} \tag{B-3}$$