TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Nachrichtentechnik

From Compact MIMO to Massive MIMO

Andrei-Stefan Nedelcu

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

genehmigten Dissertation.

Vorsitzender:	Prof. Dr.-Ing. Bernhard Seeber
Prüfer der Dissertation:	1. Prof. Dr. Gerhard Kramer
	2. Prof. Bertrand Hochwald, PhD.

Die Dissertation wurde am 16.06.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 21.09.2021 angenommen.

Preface

I am deeply grateful to many friends and family that have either directly or indirectly contributed to this thesis. First, I would like to thank my advisor Prof. Gerhard Kramer for giving me this once in a lifetime opportunity. I deeply appreciate his excellent scientific insight, patient mentorship and not least his dedication without which this thesis might never have seen the light of day.

My deepest gratitude to Luca and Fabian for being wonderful collaborators. A special place remains in my heart for my colleagues that made LNT enjoyable and memorable over the years.

This adventure wouldn't have started without the relentless support of my beloved wife and without the encouragement of my adored parents. Thank you all from the bottom of my heart!

München, April 2021 Andrei Nedelcu

Contents

Zusammenfassung

In dieser Arbeit werden Multiple-Input-Multiple-Output (MIMO)-Systeme in zweierlei Hinsicht untersucht. Informationstheoretischen Konzepte verwendet um RF-Front-Ends für kompakte Antennen-Arrays mit gekoppelten Antennen zu entwerfen, und zum anderen werden nichtlineare Vorkodierungsalgorithmen für Massive MIMO Multiuser Downlink-Kanäle mit grob quantisierten Kanaleingängen betrachtet. Für den Entwurf des HF-Frontends wird gezeigt, dass eine Klasse von verlustfreien passiven Multiport-Anpassungsnetzwerken die Informationsraten bei beliebig wählbaren Antennenabständen maximiert. Diese Lösung ist jedoch empfindlich gegenüber Störungen, die durch Fertigungsvariationen verursacht werden und gilt nur innerhalb enger Frequenzbänder um die Designfrequenz. Für MIMO systeme mit mehreren Benutzern wird die Vorkodierung sowohl für flache als auch frequenzselektive Fading-Kanäle betrachtet. Es werden mehrere Optimierungsalgorithmen für grob quantisierte Eingänge untersucht, die den quadratischen Fehler am Empfänger minimieren sollen. Die Algorithmen bieten die Möglichkeit, zwischen Leistung und Komplexität abzuwägen und arbeiten nah an den unteren Schranken des minimal erreichbaren quadratischen Fehlers und arbeiten zuverlässig. Außerdem funktionieren sie über einen großen Bereich des Signal-zu-Rausch-Verhältnisses. Darüber hinaus liegen die Informationsraten innerhalb eines kleinen, konstanten Abstands von der Kapazität unendlich genau auflösender linearer Vorkodierer. Schlussendlich können die entworfene Hardware Architektur und die Kodierungsalgorithmen in bestehende Infrastruktur integriert werden und ermöglichen einen Arbeitspunkt in der Nähe des theoretischen Optimums bei geringer Komplexität.

Abstract

This thesis studies Multiple Input Multiple Output (MIMO) systems from two points of view: information theoretic design of Radio Frequency (RF) front-ends for compact antenna arrays with mutual coupling and, on the other hand, nonlinear precoding algorithms for Massive MIMO multiuser downlink channels with coarsely quantized channel inputs. For the RF front-end design, it is shown that a class of lossless passive multiport matching networks maximizes information rates at arbitrary antenna spacing. However, the solution is sensitive to perturbations caused by process variations and is valid only within narrow fractional bandwidths around the design frequency. For the multiuser MIMO precoding problem, both flat and frequency-selective fading channels are considered. Several coordinate descent algorithms for coarsely quantized input alphabets are studied that aim to minimize a Mean Squared Error (MSE) at the receiver. The algorithms offer good performance complexity trade-offs and operate close to MSE lower bounds over a broad range of Signal to Noise Ratios (SNRs). Moreover, the information rates are within a small constant gap from the capacity of infinite resolution linear Minimum Mean Squared Error (MMSE) precoders. Finally, the proposed hardware architecture and precoding algorithms can be integrated with existing infrastructure and operate close to the ideal massive MIMO regime at low complexity.

1

Introduction

Figure 1.1 [1] shows the global explosion of monthly mobile data usage, which is orders of magnitude larger than the network load incurred by voice. The overwhelming majority of data is video streaming traffic (76%) followed by social media networks with 19%. The trend is expected to continue because of the increasing resolution and services associated with streaming, e.g., Youtube 360. Communication standards such as 5G, 802.11ax and the emerging 6G rely heavily on technologies that increase data rates at scale and low cost. In particular, MIMO and massive MIMO are cornerstone physical layer technologies that can offer the desired increase in spectral and energy efficiency.

1.1. Challenges of Compact MIMO Systems

Communication standards such as 5th Generation New Radio (5G-NR) and 802.11ax are including more antennas at both ends of the communication chain. Since mobile terminals should remain small, the antenna elements must be placed closer together. When this happens, the electromagnetic field radiated by one antenna generates currents on the antennas in its vicinity, which is called *mutual coupling.* This causes the port impedance and the radiated far field patterns to change, which may result in

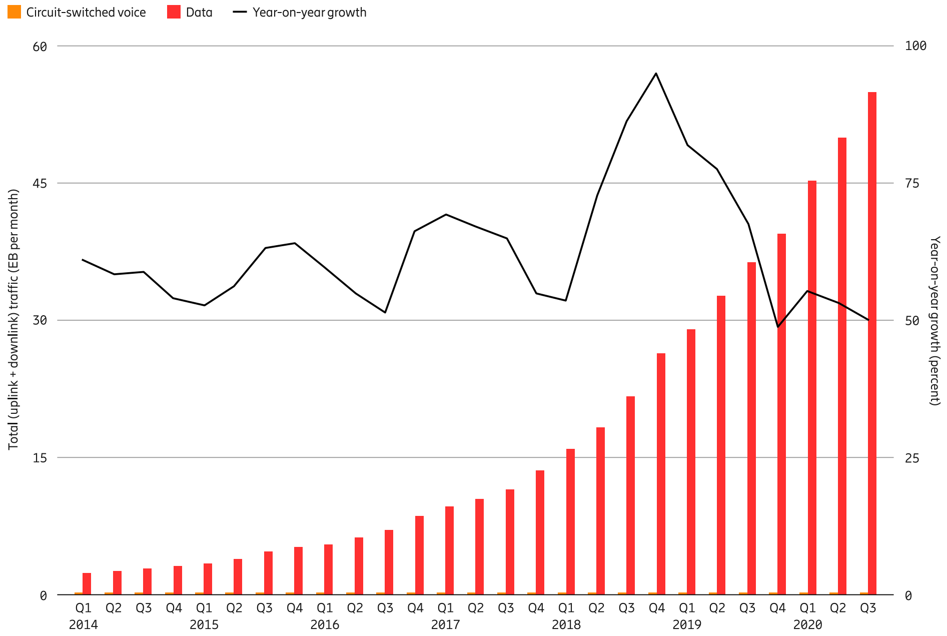


Figure 1.1.: Global mobile data traffic and year-on-year growth

power that couples back into the array, or that reflects back into the front-end due to the mismatch between the matching networks and the new antenna impedance. Moreover, the antenna currents become increasingly correlated which reduces the channel's degrees of freedom and thereby the information rates.

A secondary effect of coupling, more prevalent in the receivers, is that both signal and amplifier noise couple through the arrays. Noise coupling may seem advantageous at first, since increasing noise correlation generally results in higher rates if appropriate power allocation strategies are employed at the transmitter. However, in a physical system the noise correlation increases the total noise power because of the impedance mismatch caused by coupling.

Some of these issues were addressed in [2] that showed that internal RF front-end coupling does not impact channel capacity if the amplifier noise is ignored, and otherwise capacity is reduced. These topics were further explored in [3, 4] where matching networks were derived to minimize the noise figure and to maximize the mutual information. We provide an alternative proof of optimality and study the sensitivity of capacity to device variations. We remark that we only briefly touch

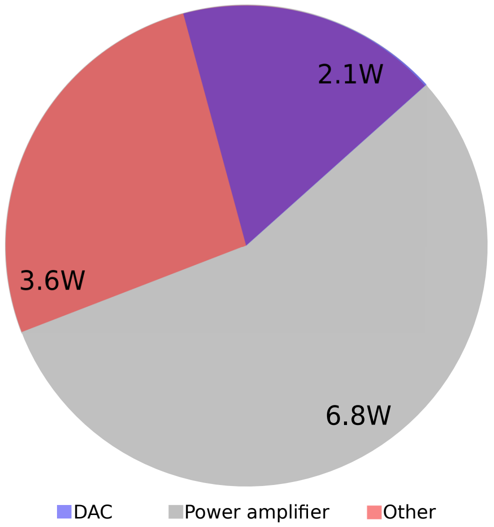


Figure 1.2.: Power consumption of a single high resolution downlink RF chain.

upon broadband matching that has been studied in more depth in [5, 6].

1.2. Challenges of Massive MIMO Systems

Massive MIMO promises unparalleled data rates and user densities, but it also requires that the base station be equipped with many more antennas than what is currently technologically feasible. Testbeds and prototypes of massive MIMO show how integration, cost and power consumption put strong limitations on scaling. To bring massive MIMO to the market one must address the following limitations, see [7].

- Power consumption due to Radio Frequency (RF) chains with a low efficiency (linear) power amplifier and a high resolution Digital to Analog Converter (DAC) and Analog to Digital Converter (ADC). Figure 1.2 breaks down the power consumption for the case of a pico-cell [7].

- Inexpensive modular hardware.

- Large physical size of antenna arrays in “low” frequency bands.

Several architectures have been proposed to address these challenges. For instance, analog beamforming [8] with and without antenna selection and hybrid analog beamforming [9] have been proposed to reduce power consumption. However, analog beamforming cannot achieve the same performance of fully digital solutions in most scenarios. In addition, hybrid architectures require increased analog complexity and high resolution phase shifters to approach what a fully digital approach can provide. To address these challenges, we design and investigate an architecture that offers algorithmic simplicity, efficient waveforms, nonlinear amplification, and low-resolution DACs.

1.3. Dissertation Overview

The dissertation has two main parts.

- ▷ Chapter 2: Information theory for coupled MIMO, including optimal matching circuits and numerical analyses of the sensitivity of information rates for narrowband coupled antenna arrays.
- ▷ Chapter 3: Efficient precoding algorithms for multiuser downlink MIMO channels with coarse discrete signaling and frequency selectivity.

1.3.1. Information Rates for Coupled MIMO

Chapter 2 studies coupled antenna arrays from an information theoretic perspective. We extend and generalize previous results in the literature and show that a class of lossless passive matching networks maximizes the mutual information and orders the channels with respect to a generalized noise figure. Our study is based on linear channels with Gaussian noise but is otherwise general with respect to the physical channel correlation structure, the antenna array, and the transmit signal covariance. We further examine the sensitivity of optimal and sub-optimal matching networks with respect to imperfections of lumped element lossless networks. We also inspect the sensitivity with respect to bandwidth, and show that the structures we used are essentially narrowband in nature.

1.3.2. Precoding for Multi-User MIMO with Discrete Signaling

Chapter 3 considers fully digital beamforming when the number of antennas at the base station is much larger than the number of (single antenna) users served in the cell. The main challenge is to develop a high rate and computationally efficient algorithm that specifies a nonlinear precoder. *High rate* means that one should approach the sum information rate of full resolution solutions such as Minimum Mean Squared Error (MMSE) or Zero Forcing (ZF) linear beamforming.

We first propose a hybrid coordinate minimization algorithm for flat fading channels. We then extend this approach to frequency-selective channels with Orthogonal Frequency Division Multiplex (OFDM). We show that the proposed algorithm has a low computational complexity as compared to state-of-the-art solutions from the literature. To evaluate information rates, we use the Generalized Mutual Information (GMI) and show that this framework fairly compares linear and nonlinear techniques. We evaluate the complexity of the proposed algorithms by developing tight upper and lower bounds on complexity based on semidefinite relaxation and branch-and-bound algorithms.

The second half of the chapter develops architectures for which our precoding algorithms are used as plug-in extensions for legacy base stations. We show that relatively simple and computationally efficient coordinate descent algorithms with offline power allocation are near-optimal for realistic wireless models.

The suggestion to study the coarsely quantized precoding problem in the GMI framework originated with my co-author Fabian Steiner [10].

1.4. Notation

We introduce notation that we use throughout the thesis.

Probability and Expectation

- ▷ The probability of event $\mathcal{A}$ is denoted by $\Pr(\mathcal{A})$.

- ▷ Uppercase letters denote discrete or continuous random variables, and lowercase letters their realizations. The density of a real, Gaussian, random variable X with mean m and variance σ^2 is written as $\mathcal{N}(m, \sigma^2)$, and we write $X \sim \mathcal{N}(m, \sigma^2)$. The density of a circularly-symmetric, complex, Gaussian, random variable X with variance $\mathrm{E}\left[|X|^2\right] = \sigma^2$ is written as $\mathcal{CN}(0, \sigma^2)$. Similarly, the density of $m + X$ is $\mathcal{CN}(m, \sigma^2)$.

- ▷ The expectation of a random variable Y is written as $\mathrm{E}\left[Y\right]$ and the expectation of Y conditioned on the random variable Z is written as $\mathrm{E}\left[Y|Z\right]$. The expectation of Y conditioned on the event $Z = z$ is written as $\mathrm{E}\left[Y|Z = z\right]$.

- ▷ We denote a scalar random variable as X and a random column vector as $\boldsymbol{X}$. The covariance matrix of $\boldsymbol{Z}$ is $\boldsymbol{C_Z} = \mathrm{E}\left[(\boldsymbol{Z} - \mathrm{E}\left[\boldsymbol{Z}\right])(\boldsymbol{Z} - \mathrm{E}\left[\boldsymbol{Z}\right])^{\mathrm{H}}\right]$. The density of a circularly-symmetric, complex, Gaussian, random vector $\boldsymbol{Z}$ is written as $\mathcal{CN}(\boldsymbol{0}, \boldsymbol{C_Z})$. Similarly, the density of $\boldsymbol{m} + \boldsymbol{Z}$ is $\mathcal{CN}(\boldsymbol{m}, \boldsymbol{C_Z})$.

Information Measures

- ▷ For a pair of random variables X and Y with joint density $p_{XY} = p_X p_{Y|X}$, the mutual information of X and Y is

$$\mathrm{I}(X;Y) = \mathrm{E}\left[\log_2\left(\frac{p_{Y|X}(Y|X)}{p_Y(Y)}\right)\right] \tag{1.1}$$

where p_Y is the density of Y. The same definition is used if X has a discrete alphabet.

Sets, Vectors, Matrices and Norms

- ▷ We write the set of integers from 1 to M as $\mathbb{U} := \{1, \cdots, M\}$. The Cartesian product of a set $\mathcal{A}$ with itself is denoted as $\mathcal{A}^2$.

- ▷ Vectors are denoted in bold lowercase letters $\boldsymbol{x}$ and matrices are denoted with bold uppercase letters $\boldsymbol{X}$. The distinction between random vectors and fixed-value matrices will be clear from the context. The transpose and Hermitian transpose of a vector $\boldsymbol{z}$ are given by $\boldsymbol{z}^T$ and $\boldsymbol{z}^H$, respectively. The real an imaginary parts of a complex valued vector $\boldsymbol{z}$ are given by $\Re\{\boldsymbol{z}\}$ and $\Im\{\boldsymbol{z}\}$.

- ▷ We denote the trace of a matrix $\boldsymbol{C}$ as $\operatorname{Tr}(\boldsymbol{C})$ and the determinant as $\det(\boldsymbol{C})$. The inverse of a matrix is denoted as $\boldsymbol{C}^{-1}$.

- ▷ We write $\boldsymbol{A} \succeq \boldsymbol{B}$ if the matrix $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite and $\boldsymbol{A} \succ \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B}$ is positive definite. Recall that the $n \times n$ matrix $\boldsymbol{A}$ is positive semidefinite if and only if $\boldsymbol{z}^H \boldsymbol{A} \boldsymbol{z} \geq 0$ for all $\boldsymbol{z} \in \mathbb{C}^n$.

- ▷ We write the p-norm of a vector as $\|\boldsymbol{u}\|_p$ and the squared p-norm as $\|\boldsymbol{u}\|_p^2$. The *infinity*-norm is denoted as $\|\boldsymbol{u}\|_\infty$ and is defined as the maximum of the absolute values of $\boldsymbol{u}$.

- ▷ The Hadamard (element-wise) product of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ is denoted as $\boldsymbol{A} \circ \boldsymbol{B}$.

Complexity

- ▷ We use the Bachmann–Landau big-O notation $f(x) = \mathcal{O}(g(x))$ to say that $f(x)$ does not grow faster than the positive-valued $g(x)$ as a function of x, i.e., there are finite M and x_0 such that $|f(x)| \leq Mg(x)$ for all $x \geq x_0$.

2

Information Rates for Coupled MIMO

2.1. Introduction

Antenna arrays should be made compact to save space but antenna proximity causes coupling which may impact performance. Matching circuits placed at the transmitter and receiver antennas serve to de-couple the antennas, or even better to maximize the mutual information between the information bits and the received signal.

We investigate matching circuits for narrowband signals. Sec. 2.2 reviews models and theory for optimal matching for Single Input Single Output (SISO) systems. Sec. 2.3 and Sec. 2.4 review models and theory for Multiple Input Multiple Output (MIMO) systems, including the capacity for passive, lossless, and reciprocal matching circuits. Sec. 2.5 presents a sensitivity analysis, where the sensitivity is computed by varying the antenna spacing, the device parameters, and the bandwidth. Sec. 2.6 concludes the chapter.

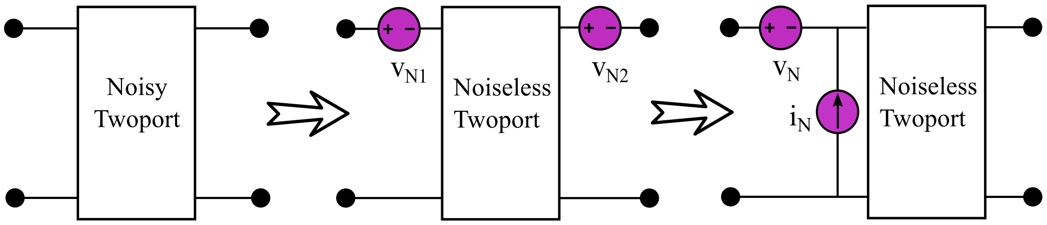


Figure 2.1.: Noisy fourpole models and equivalent circuits

2.2. SISO Antenna Systems

2.2.1. Amplifier Noise

Rothe and Dahlke [11] introduced a theory to characterize the noise behaviour of passive and active fourpoles. The equivalent Thévenin representation of a noisy fourpole is shown in Fig. 2.1. The internal noise sources can be equivalently represented by a noiseless twoport and external noise voltage generators at the input, output or both. The noise sources represent physical phenomena in a small volume, and they are correlated in general. The output noise voltage source can be represented by an equivalent input noise current that may be correlated with the input noise voltage [11]. This representation separates the effects of noise from the noise-free downstream circuits. Furthermore, the signal to noise ratio (SNR) is computed conveniently because the noise and signal are known at the same ports in the equivalent circuit.

We now generally follow the notation in [12], except that we write random variables with uppercase letters, e.g., V_N and I_N, and their realizations with the corresponding lowercase letters, e.g., v_n and i_N. Define the quantities:

$$\beta = \mathrm{E}\left[|I_N|^2\right] = kT_0 W G_N \tag{2.1}$$

$$R_N = \sqrt{\frac{\mathrm{E}\left[|V_N|^2\right]}{\mathrm{E}\left[|I_N|^2\right]}} \tag{2.2}$$

$$\rho = \frac{\mathrm{E}\left[V_N I_N^\star\right]}{\sqrt{\mathrm{E}\left[|I_N|^2\right]\mathrm{E}\left[|V_N|^2\right]}} \tag{2.3}$$

where β is the input-referenced noise current total power, k is Boltzmann's constant, T_0 is the environment equilibrium temperature, W is the bandwidth, G_N is the equivalent noise conductance of the amplifier, R_N is the equivalent noise resistance, and ρ is the correlation coefficient of the noise voltage and the noise current. The noise random variables I_N and V_N are modeled as zero mean Gaussian with variances β and βR_N^2 respectively. This is consistent with the definitions of the noise parameters from [11] where the noise voltage $V_N = V_{Nuncorr} + Z_{corr} I_N$ is separated into a correlated $Z_{corr} I_N$ and an uncorrelated $V_{Nuncorr}$ component.

2.2.2. Antenna Noise

Apart from the noise originating at the active elements (amplifiers) there is also noise at the antenna [13]. Using the Rayleigh-Jeans approximation (which approximates the Planck black body radiation law for the range of frequencies at which the noise power spectral density (Power Spectral Density (PSD)) is white), the antenna noise source can be represented by an equivalent Thévenin voltage source $V_{SN} = v_{SN}$ with [13]:

$$\mathrm{E}\left[|V_{SN}|^2\right] = 4kT_{ant}WR_{AR} \tag{2.4}$$

where the antenna noise temperature T_{ant} is the equivalent temperature of a resistor with resistance R_{AR} required to produce the same noise power as the actual environment seen by the antenna.

Antenna noise may be generated by the surrounding environment or by loss mechanisms of the antenna itself. The background noise temperature of the antenna is given by integrating over the background noise temperature weighted by the antenna directivity [13]:

$$T_{br} = \frac{\int_{\phi=0}^{2\pi} \int_{\theta=0}^{\pi} T_B(\theta, \phi) D(\theta, \phi) \, \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi}{\int_{\phi=0}^{2\pi} \int_{\theta=0}^{\pi} D(\theta, \phi) \, \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi} \tag{2.5}$$

where T_{br} is the *brightness temperature* or the equivalent total noise temperature of the antenna absorbed from background sources, $T_B(\theta, \phi)$ is the distribution of the background temperature, $D(\theta, \phi)$ is the directivity of the antenna, and θ and ϕ are the elevation and azimuth angles, respectively. The directivity is used here instead of gain such that the brightness temperature captures only the influence of the background

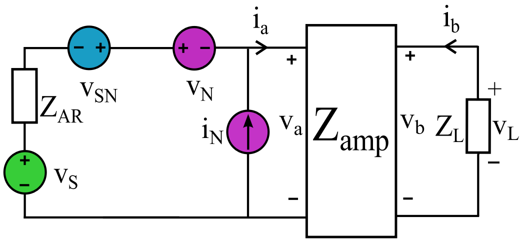


Figure 2.2.: Single antenna front-end receiver without impedance matching

and not also the losses of the antenna.

If the antenna is lossy then it can be modeled by an equivalent lossless antenna followed by an attenuator with loss factor $L = 1/\eta_{rad}$ where η_{rad} is the radiation efficiency. The overall noise temperature of a lossless antenna cascaded with an attenuator at temperature T_p is [13]:

$$T_{ant} = \frac{1}{L} T_{br} + \frac{L-1}{L} T_p \, . \tag{2.6}$$

As in [13] the antenna noise temperature is referenced at the antenna terminals. If the antenna is terminated on a matched load then the total noise power delivered by the antenna is $kT_{ant}W$ Watts.

2.2.3. Circuit Model

Consider the circuit in Fig. 2.2. The information-carrying voltage source is v_S, the antenna noise source is v_{SN}, and the antenna impedance is Z_{AR} with real part $\Re\{Z_{AR}\} = R_{AR}$. We model the voltages as random variables with zero mean. The noisy two-port has the impedance matrix

$$\boldsymbol{Z}_{amp} = \begin{bmatrix} Z_{amp11} & Z_{amp12} \\ Z_{amp21} & Z_{amp22} \end{bmatrix} \tag{2.7}$$

and the currents and voltages at the ports are related by:

$$\begin{bmatrix} v_a \\ v_b \end{bmatrix} = \boldsymbol{Z}_{amp} \begin{bmatrix} i_a \\ i_b \end{bmatrix} . \tag{2.8}$$

With the conventions for voltage and currents (voltage drop from the + pole to the - pole and current flowing into the + pole and out of the - pole with positive signs) we have $v_b = v_L = -i_b Z_L$. Applying (2.8), we can write

$$-Z_{amp21} i_a = (Z_{amp22} + Z_L) i_b \tag{2.9}$$

and using Kirchhoff's laws we compute

$$\begin{align} 0 &= -v_S - v_{SN} + v_N + Z_{AR}(i_a - i_N) + i_a Z_{amp11} + i_b Z_{amp12} \tag{2.10} \\ 0 &= (Z_{amp22} + Z_L) i_b + Z_{amp21} i_a. \tag{2.11} \end{align}$$

Solving for i_b, we have

$$\begin{align} & i_b((Z_{amp22} + Z_L)(Z_{amp11} + Z_{AR}) - Z_{amp21} Z_{amp12}) \\ & \quad = (-v_S - v_{SN} + v_N - Z_{AR} i_N) Z_{amp21} \tag{2.12} \end{align}$$

which together with $v_b = v_L = -i_b Z_L$ gives

$$v_L = \frac{Z_{amp21} Z_L (v_S + v_{SN} - v_N + Z_{AR} i_N)}{(Z_{amp22} + Z_L)(Z_{amp11} + Z_{AR}) - Z_{amp21} Z_{amp12}}. \tag{2.13}$$

Define the transfer function

$$T = \frac{Z_{amp21} Z_L}{(Z_{amp22} + Z_L)(Z_{amp11} + Z_{AR}) - Z_{amp21} Z_{amp12}} \tag{2.14}$$

and $n_N = v_{SN} - v_N + Z_{AR} i_N$ so that we have the noisy channel output

$$v_L = T(v_S + n_N). \tag{2.15}$$

2.2.4. Noise Figure

A traditional quantity in our context is the *noise figure*

$$F = 1 + \frac{T_{ant}}{T_0} \left(\frac{\mathrm{E}\left[|N_N|^2\right]}{\mathrm{E}\left[|V_{SN}|^2\right]} - 1 \right)$$

$$= 1 + \frac{\beta R_N^2 - 2\Re\{Z_{AR}^{\star}\rho\beta R_N\} + \beta|Z_{AR}|^2}{4kT_0W\Re\{Z_{AR}\}} \tag{2.16}$$

where the second step follows by

$$\begin{aligned} \mathrm{E}\left[|N_N|^2\right] &= \mathrm{E}\left[|(V_{SN} - V_N + Z_{AR}I_N)|\right]^2 \\ &= \mathrm{E}\left[|V_{SN}|^2\right] + \beta R_N^2 - 2\Re\{Z_{AR}^{\star}\rho\beta R_N\} + \beta|Z_{AR}|^2. \end{aligned} \tag{2.17}$$

The background noise V_{SN} captured by the antenna is assumed to be independent of the amplifier noise V_N and I_N because of the separate random mechanisms for these noise sources. Note that F is computed at a source noise temperature T_0.

We consider only passive antennas that have $\Re\{Z_{AR}\} > 0$, which means that their radiation efficiency is non-negative and nonzero. Suppose we could optimize Z_{AR} to minimize F. We compute

$$\begin{aligned} \frac{\partial F}{\partial\Im\{Z_{AR}\}} &= \frac{\beta}{2kT_0W\Re\{Z_{AR}\}}\left(-\Im\{\rho\}R_N + \Im\{Z_{AR}\}\right) = 0 \\ \frac{\partial F}{\partial\Re\{Z_{AR}\}} &= \frac{\beta}{4kT_0W\Re\{Z_{AR}\}^2}\left(\Re\{Z_{AR}\}^2 - \Im\{Z_{AR}\}^2 - R_N^2 + 2\Im\{Z_{AR}\}R_N\Im\{\rho\}\right) = 0 \end{aligned}$$

and the Z_{AR} that minimizes the noise figure is

$$Z_{opt} = R_N\left(\sqrt{1 - \Im\{\rho\}^2} + j\Im\{\rho\}\right). \tag{2.18}$$

The resulting minimum noise figure after substituting (2.18) in (2.16) is

$$F_{min} = 1 + \frac{2\beta R_N}{4kT_0W}\left(\sqrt{1 - \Im\{\rho\}^2} - \Re\{\rho\}\right) \tag{2.19}$$

and one may write

$$F = F_{min} + \frac{\beta}{4kT_0W\Re\{Z_{AR}\}}|Z_{AR} - Z_{opt}|^2. \tag{2.20}$$

We add a few remarks [14].

▷ The noise parameter R_N depends on the technology. For Complementary Metal

Oxide Semiconductor (CMOS) R_N is relatively stable across different instances of the manufacturing process.

- ▷ R_N may vary with the impedance Z_{AR}, in which case the optimization changes.
- ▷ Noise matching minimizes F but creates a mismatch from the input impedance that maximizes the amplifier gain. Noise matching and maximum power transfer matching are often performed jointly as they trade-off against each other.
- ▷ Further design steps (negative feedback) can insure that the input impedance of the amplifier approaches $Z^{\star}_{opt}$, a technique denoted as simultaneous noise and input matching. This technique ensures that the optimum noise behavior is achieved and the gain of the amplifier is optimized. The authors in [14] observe that the gain behavior is much less sensitive to parameter variations than F_{min} and R_N.
- ▷ The quality of the matching depends on the fabrication technology and the size and power dissipation of the transistors. Perfect noise matching may not be possible in practice.

2.2.5. Impedance Matching

We next derive a circuit to transform the antenna impedance to the desired Z_{opt}. Consider a single frequency optimization at or very close to the antenna resonance, which in most cases is also the carrier frequency. A *lossless reciprocal* matching circuit is shown in Fig. 2.3. The lossless constraint means that the matching circuit has pure reactive elements (impedance has only an imaginary part) and does not increase the antenna temperature.

A two-port lossless reciprocal matching circuit in matrix form can be expressed as

$$Z_M = \begin{bmatrix} Z_{M11} & Z_{M12} \\ Z_{M12} & Z_{M22} \end{bmatrix} = j \begin{bmatrix} X_{M11} & X_{M12} \\ X_{M12} & X_{M22} \end{bmatrix} \tag{2.21}$$

where the X_{M11}, X_{M12}, X_{M22} are real. In other words, the lossless restriction requires Z_M to have purely imaginary entries, and the reciprocal constraint means $Z_M = Z_M^{\mathrm{T}}$.

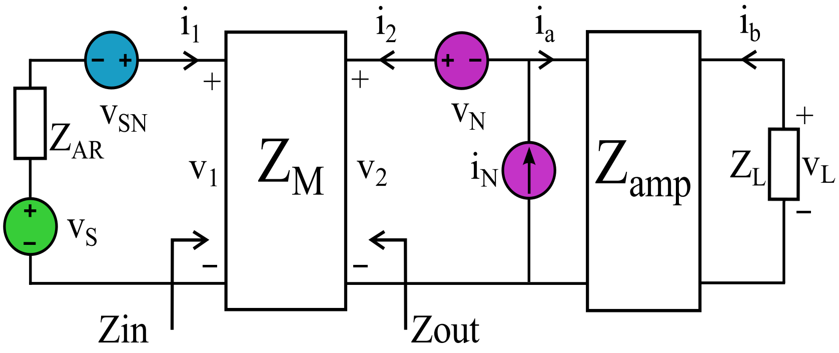


Figure 2.3.: Single Antenna Front-End Receiver with Impedance Matching

Suppose we wish to make the impedance Z_{out} at the output of the matching network to be Z_{opt}:

$$Z_{out} = -\frac{Z_{M12}Z_{M12}}{Z_{AR}+Z_{M11}} + Z_{M22} = R_N\left(\sqrt{1-\Im\{\rho\}^2} - j\Im\{\rho\}\right). \tag{2.22}$$

To accomplish this, one can choose several values of the elements of the matching network to satisfy (2.22). For example, if Z_{M11} is a free parameter then Z_{M22} depends on this choice and the other fixed values. If we choose $Z_{M11} = -j\Im\{Z_{AR}\}$ we obtain

$$Z_M = \begin{bmatrix} -j\Im\{Z_{AR}\} & \pm j\sqrt{\Re\{Z_{AR}\}\Re\{Z_{opt}\}} \\ \pm j\sqrt{\Re\{Z_{AR}\}\Re\{Z_{opt}\}} & j\Im\{Z_{opt}\} \end{bmatrix}. \tag{2.23}$$

On the other hand, if the input impedance of the amplifier is $Z^{\star}_{opt}$ then the input impedance of the matching network is:

$$Z_{in} = -\frac{Z_{M12}Z_{M12}}{Z^{\star}_{opt}+Z_{M22}} + Z_{M11}. \tag{2.24}$$

If $Z_{in} = Z^{\star}_{AR}$ then we have an input conjugate match and the matching maximizes power transfer to the amplifier.

The above considerations for optimization confirm practices from microwave antenna design [13] [15]. These guidelines are often derived using the scattering parameter formalism which is equivalent to the Z formalism used here.

2.2.6. Source-Load Information Rates

Assuming $T \neq 0$, the mutual information between the source and the load is:

$$\begin{aligned} I(V_S;V_L) &= I(V_S;T^{-1}V_L) \\ &= h(V_S+N_N) - \log_2\left(\pi e \operatorname{E}\left[|N_N|^2\right]\right) \\ &\overset{(a)}{\leq} \log_2\left(\pi e(\operatorname{E}\left[|V_S|^2\right] + \operatorname{E}\left[|N_N|^2\right]\right) - \log_2\left(\pi e \operatorname{E}\left[|N_N|^2\right]\right) \\ &\overset{(b)}{=} \log_2\left(1 + \frac{P_{sig}}{(F-1+T_{ant}/T_0)4kT_0W\Re\{Z_{AR}\}}\right) \end{aligned} \tag{2.25}$$

where (a) follows by a maximum entropy theorem [16](Chapter 12, pg. 409), and (b) by using $P_{sig} = \operatorname{E}[|V_S|^2]$ and the definition of F.

Note that F is the SNR drop across the receiver because of the intrinsic added noise of the amplifier with T_0 as reference noise temperature for the source. Note also that the SNR drop across the receiver is computed for a source at noise temperature T_{ant} whereas F is computed at T_0. This difference was stressed in [13, p. 496].

A natural question is why an amplifier is needed, since it reduces the mutual information. The reason is that for digital processing the electric signal generated at the antenna ports must be sampled and quantized. Quantization requires adapting to the dynamic range of the ADC which is often on the order of volts as opposed to fractions of microvolts at the antenna terminals in most cellular communications.

In fact, multiple stages of amplification, mixing, filtering usually occur, depending on the receiver architecture [17], see Fig. 2.4. The cascaded system noise figure is [13]

$$F_{receiver} = F_{LNA} + \frac{F_2-1}{G_{LNA}} + \frac{F_3-1}{G_2G_{LNA}} + ... \tag{2.26}$$

where F_{LNA} is the noise figure of the low noise amplifier, G_{LNA} is the transducer gain of the Low Noise Amplifier (LNA), $F_2, F_3, ...$ are the noise figures of further stages of the receiver and $G_2, G_3, ...$ are the transducer gains of these subsequent stages. We will assume the gain of the LNA is large enough so the noise contributions of the stages following the LNA do not significantly change the receiver noise figure.

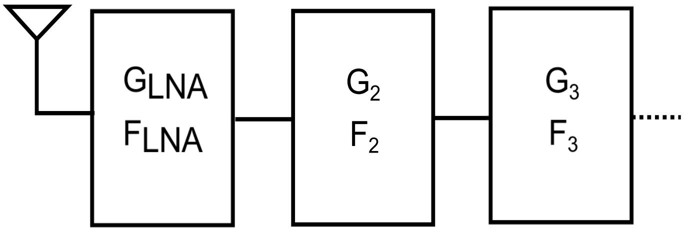


Figure 2.4.: Receiver Chain

2.3. MIMO Channels

2.3.1. Input-Output Information Rates

The system model of Sec. 2.2 can be extended to MIMO systems (see, e.g., [4, 12]). Consider the MIMO flat fading channel

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{z} \tag{2.27}$$

where $\boldsymbol{x} \in \mathbb{C}^M$ is the transmitted signal, $\boldsymbol{y} \in \mathbb{C}^N$ is the received signal, $\boldsymbol{H} \in \mathbb{C}^{M \times N}$ is the channel transfer matrix and $\boldsymbol{z} \in \mathbb{C}^N$ is AWGN noise. The total average transmit power and the transmit covariance matrix are given by

$$P_{\boldsymbol{X}} = \mathrm{Tr}\left(\boldsymbol{C}_{\boldsymbol{X}}\right), \quad \boldsymbol{C}_{\boldsymbol{X}} = \mathrm{E}\left[\boldsymbol{X}\boldsymbol{X}^{\mathrm{H}}\right] \tag{2.28}$$

respectively, where $\boldsymbol{X}$ is assumed to have zero mean. The mutual information when the receiver has perfect channel state information (Channel State Information (CSI)) is

$$I(\boldsymbol{X};\boldsymbol{Y}) = h(\boldsymbol{Y}) - h(\boldsymbol{Y}|\boldsymbol{X}) \tag{2.29}$$

where $h(\boldsymbol{Y})$ denotes the differential entropy of $\boldsymbol{Y}$ and $h(\boldsymbol{Y}|\boldsymbol{X})$ is the average differential entropy of $\boldsymbol{Y}$ conditioned on $\boldsymbol{X}$. Since $\boldsymbol{X}$ and $\boldsymbol{N}$ are independent, (2.29) becomes

$$I(\boldsymbol{X};\boldsymbol{Y}) = h(\boldsymbol{Y}) - h(\boldsymbol{Z}). \tag{2.30}$$

The differential entropy of circularly symmetric complex Gaussian noise is be expressed in terms of its covariance matrix [18] as

$$h(\boldsymbol{Y}) = \log_2 \det(\pi e\, \boldsymbol{C}_{\boldsymbol{Y}}). \tag{2.31}$$

Using $\boldsymbol{C}_{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{C}_{\boldsymbol{X}}\boldsymbol{H}^{\mathrm{H}} + \boldsymbol{C}_{\boldsymbol{Z}}$ gives

$$\begin{aligned} I(\boldsymbol{X};\boldsymbol{Y}) &= \log_2 \det\left(\pi e(\boldsymbol{H}\boldsymbol{C}_{\boldsymbol{X}}\boldsymbol{H}^{\mathrm{H}} + \boldsymbol{C}_{\boldsymbol{Z}})\right) - \log_2 \det(\pi e \boldsymbol{C}_{\boldsymbol{Z}}) \\ &= \log_2 \det\left(\boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{Z}}^{-1}\boldsymbol{H}\boldsymbol{C}_{\boldsymbol{X}}\boldsymbol{H}^{\mathrm{H}}\right). \end{aligned} \tag{2.32}$$

Suppose the transmitter does not know $\boldsymbol{H}$ and chooses to distribute power equally across all signal dimensions, which is known to be the optimum strategy with Gaussian signaling if the channel is random and has certain symmetries [18]. The covariance matrix of the channel input is thus given by

$$\boldsymbol{C}_{\boldsymbol{X}} = \frac{P_{\boldsymbol{X}}}{M}\boldsymbol{I}. \tag{2.33}$$

If we assume spatially uncorrelated Gaussian noise at the receiver, then we have

$$\boldsymbol{C}_{\boldsymbol{Z}} = \mathrm{E}\left[\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{H}}\right] = P_{\boldsymbol{Z}}\boldsymbol{I}. \tag{2.34}$$

Inserting the expressions (2.33) and (2.34) into (2.32) gives

$$I(\boldsymbol{X};\boldsymbol{Y}) = \log_2 \det\left(\boldsymbol{I} + \frac{\mathrm{SNR}}{M}\,\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}}\right) \tag{2.35}$$

with the SISO SNR $= P_{\boldsymbol{X}}/P_{\boldsymbol{Z}}$ and with units bit/s/Hz or bpcu.

2.3.2. Ergodic Information Rates

A spatially white channel model is defined by the random matrix $\boldsymbol{H}_w$ with

$$\mathrm{E}\left[[\boldsymbol{H}_w]_{i,j}\right] = 0 \tag{2.36a}$$

$$\mathrm{E}\left[|[\boldsymbol{H}_w]_{i,j}|^2\right] = 1 \tag{2.36b}$$

$$\mathrm{E}\left[[\boldsymbol{H}_w]_{i,j}[\boldsymbol{H}_w]^*_{m,n}\right] = 0 \text{ if } i \neq m \text{ or } j \neq n \tag{2.36c}$$

$$\mathrm{E}\left[[\boldsymbol{H}_w]_{i,j}[\boldsymbol{H}_w]_{m,n}\right] = 0 \text{ for all } i, j, m, n. \tag{2.36d}$$

The instantaneous information rates are random variables that depend on the realizations of $\boldsymbol{H}_w$. The ergodic channel capacity is defined as the expected instantaneous information rate, and can generally be determined via Monte Carlo simulations by sampling $\boldsymbol{H}_w$ and averaging:

$$\begin{aligned} C_{\text{erg}} &= \mathrm{E}\left[\log_2 \det\left(\boldsymbol{I} + \frac{\mathrm{SNR}}{M}\boldsymbol{H}_w\boldsymbol{H}_w^{\mathrm{H}}\right)\right] \\ &\approx \frac{1}{S}\sum_{i=1}^{S}\log_2 \det\left(\boldsymbol{I} + \frac{\mathrm{SNR}}{M}\boldsymbol{H}_i\boldsymbol{H}_i^{\mathrm{H}}\right) \end{aligned} \tag{2.37}$$

where $\boldsymbol{H}_i$ is the i-th realization of $\boldsymbol{H}_w$. The accuracy of the ergodic channel capacity estimate depends on the number of experiments. In this chapter, we use sample sizes of at least $S = 50,000$ independent channel realizations.

2.3.3. Kronecker Fading Model

The Kronecker channel model [19] captures the influence of the relative positions of antennas and the incident field angle of arrival distribution on the correlation between the signals at the output ports of the antennas. The model separates the spatial correlation at the receiver and the transmitter sides, i.e., it essentially assumes that correlation is a local effect given by the geometry in the proximity of the antenna arrays and that the rest of the macroscopic channel is a rich multipath environment. A Rayleigh fading channel with double-sided correlation according to the Kronecker model is given by

$$\boldsymbol{H} = \boldsymbol{R}_r^{\frac{1}{2}}\boldsymbol{H}_w\boldsymbol{R}_t^{\frac{1}{2}} \tag{2.38}$$

where $\boldsymbol{R}_r$ and $\boldsymbol{R}_t$ are positive definite Hermitian matrices that specify the receive and transmit correlations. The spatially white channel $\boldsymbol{H}_w$ has the defined properties introduced in (2.36a)-(2.36d). The correlation matrices are normalized such that the $\mathrm{E}\left[\mathrm{Tr}(\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}})\right] = MN$. The off-diagonal elements are complex values that specify the spatial correlation.

The average mutual information with Gaussian inputs is now

$$\begin{aligned} C_{\text{erg}} &= \mathrm{E}\left[\log_2 \det\left(\boldsymbol{I} + \frac{\text{SNR}}{N}\,\boldsymbol{H}\boldsymbol{H}^{\text{H}}\right)\right] \\ &= \mathrm{E}\left[\log_2 \det\left(\boldsymbol{I} + \frac{\text{SNR}}{N}\,\boldsymbol{R}_r\boldsymbol{H}_w\boldsymbol{R}_t\boldsymbol{H}_w^{\text{H}}\right)\right] \end{aligned} \tag{2.39}$$

where we have used Sylvester's determinant identity

$$\det\left(\boldsymbol{I} + \boldsymbol{A}\boldsymbol{B}\right) = \det\left(\boldsymbol{I} + \boldsymbol{B}\boldsymbol{A}\right) \tag{2.40}$$

if the dimensions of $\boldsymbol{A}$ and $\boldsymbol{B}$ are compatible. We simulate the ergodic channel capacity of a 2×2 MIMO system and examine the resulting capacity for different correlation coefficients at the transmitter and receiver as depicted in Fig. 2.5. The correlation matrices for the 2×2 MIMO system are

$$\boldsymbol{R}_{t/r} = \begin{bmatrix} 1 & \rho_{t/r}^* \\ \rho_{t/r} & 1 \end{bmatrix}$$

where $\rho_{t/r}$ is the correlation coefficient for either the transmitter or receiver. We see that spatial correlation decreases the channel capacity. If we are interested in the channel capacity at high SNR, then we use (2.39) and neglect the identity matrix resulting in:

$$\begin{aligned} C_{\text{erg}} &\approx \mathrm{E}\left[\log_2 \det\left(\frac{\text{SNR}}{M}\,\boldsymbol{R}_r\boldsymbol{H}_w\boldsymbol{R}_t\boldsymbol{H}_w^{\text{H}}\right)\right] \\ &= \mathrm{E}\left[\log_2 \det\left(\frac{\text{SNR}}{M}\,\boldsymbol{H}_w\boldsymbol{R}_t\boldsymbol{H}_w^{\text{H}}\right)\right] + \log_2 \det\left(\boldsymbol{R}_r\right) \\ &= \underbrace{\mathrm{E}\left[\log_2 \det\left(\frac{\text{SNR}}{M}\,\boldsymbol{H}_w^{\text{H}}\boldsymbol{H}_w\right)\right]}_{\approx\, C_{\text{erg, no spatial correlation}}} + \log_2 \det\left(\boldsymbol{R}_r\right) + \log_2 \det\left(\boldsymbol{R}_t\right). \end{aligned} \tag{2.41}$$

The correlation matrix $\boldsymbol{R}_{t/r}$ is a positive Hermitian matrix, therefore all its eigenvalues satisfy $\lambda_i(\boldsymbol{R}_{t/r}) > 0$. On the other hand, the determinant is upper bounded $\det(\boldsymbol{R}_{t/r}) \leq 1$. Hence we have $\log_2(\det(\boldsymbol{R}_{t/r})) \leq 0$. Observe that the capacity is only

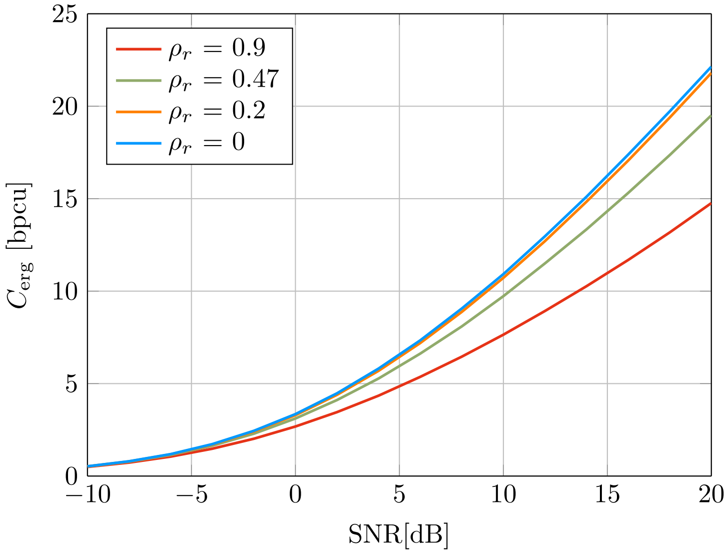


Figure 2.5.: Ergodic rates (2.39) for a 2×2 MIMO system with $\rho_t = 0.15$ and different correlation coefficients ρ_r at the receiver

reduced by correlations at either the transmitter or receiver. Moreover, in the limit of high SNR the reduction is determined only by the determinants of the correlation matrices that do not depend on SNR and therefore do not affect the growth rate of capacity with SNR.

2.4. MIMO Antenna Systems

Consider the MIMO radio model developed in [2,12,20]. Fig. 2.6 shows a system with M transmit antennas, N receive antennas, N amplifiers, a $2M \times 2M$ matching circuit at the transmitter, and a $2N \times 2N$ matching circuit at the receiver (more generally, one could use $A \times B$ matching circuits with $A \neq B$). Our focus will be on matching and we begin with a narrow-band assumption, i.e., the bandwidth is a small fraction of the carrier frequency. We will assume as in Sec. 2.2 that the matching networks are passive, lossless, and reciprocal. We consider amplifiers operating in the linear regime.

We continue to focus on the case $M = N$. At the transmitter the RF input voltages

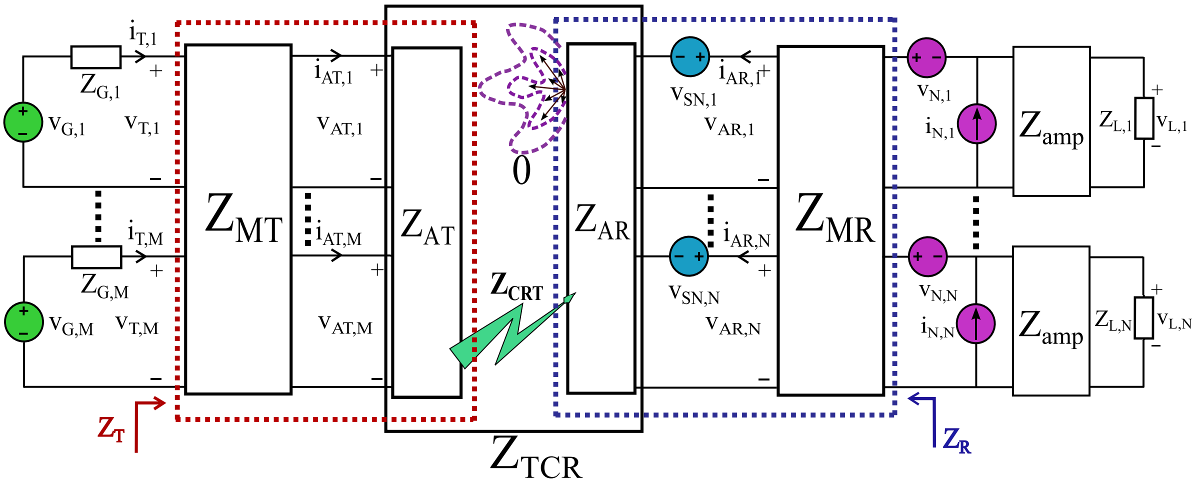


Figure 2.6.: MIMO front-end transceiver with matching

are collected into a vector:

$$\boldsymbol{v}_G = \left[v_{\mathrm{G},1}\ v_{\mathrm{G},2}\ \dots\ v_{\mathrm{G},N} \right]^T . \tag{2.42}$$

The voltage $\boldsymbol{v}_T$ at the input of the transmitter matching network is thus

$$\boldsymbol{v}_T = \boldsymbol{v}_G - \boldsymbol{Z}_G \boldsymbol{i}_T \tag{2.43}$$

where $\boldsymbol{v}_G$ is the source voltage, $\boldsymbol{Z}_G$ is the source impedance, and $\boldsymbol{i}_T$ is the loop current. The input-output relation of the impedance matching network $\boldsymbol{Z}_{\mathrm{MT}}$ is given by

$$\begin{bmatrix} \boldsymbol{v}_T \\ \boldsymbol{v}_{AT} \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{Z}_{MT11} & \boldsymbol{Z}_{MT12} \\ \boldsymbol{Z}_{MT21} & \boldsymbol{Z}_{MT22} \end{bmatrix}}_{\boldsymbol{Z}_{MT}} \begin{bmatrix} \boldsymbol{i}_T \\ -\boldsymbol{i}_{AT} \end{bmatrix} . \tag{2.44}$$

Similarly, we have

$$\begin{bmatrix} \boldsymbol{v}_{AT} \\ \boldsymbol{v}_{AR} \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{Z}_{AT} & \boldsymbol{Z}_{ATR} \\ \boldsymbol{Z}_{ART} & \boldsymbol{Z}_{AR} \end{bmatrix}}_{\boldsymbol{Z}_{TCR}} \begin{bmatrix} \boldsymbol{i}_{AT} \\ \boldsymbol{i}_{AR} \end{bmatrix} \tag{2.45}$$

and we assume that the voltage at the input of the receive array does not couple back into the transmit array, i.e., back-scattering is negligible and $\boldsymbol{Z}_{ATR} = \boldsymbol{0}$.

For instance, at the antenna array port we have

$$\boldsymbol{v}_{AT} = \boldsymbol{Z}_{AT}\boldsymbol{i}_{AT}. \tag{2.46}$$

Substituting (2.46) in the second equation of (2.44) and rearranging for $\boldsymbol{i}_{AT}$ gives

$$\boldsymbol{i}_{AT} = \left(\boldsymbol{Z}_{AT} + \boldsymbol{Z}_{M22}\right)^{-1} \boldsymbol{Z}_{\mathrm{MT21}}\boldsymbol{i}_{T}. \tag{2.47}$$

Now use the first row of (2.44) and (2.47) to write

$$\boldsymbol{v}_{T} = \underbrace{\Big(\boldsymbol{Z}_{\mathrm{MT11}} - \boldsymbol{Z}_{\mathrm{MT12}}(\boldsymbol{Z}_{\mathrm{AT}} + \boldsymbol{Z}_{\mathrm{MT22}})^{-1}\boldsymbol{Z}_{\mathrm{MT21}}\Big)}_{\boldsymbol{Z}_{\mathrm{T}}}\boldsymbol{i}_{T} \tag{2.48}$$

where $\boldsymbol{Z}_{\mathrm{T}}$ is the transformed input impedance matrix when looking into the impedance matching circuit.

The physical channel $\boldsymbol{Z}_{RT}$ can be modeled by methods presented in [21]. We again use the Kronecker model [21] and compute the following source-load voltage equation (see [12, eq. (16)]):

$$\boldsymbol{v}_{L} = \boldsymbol{C}_{L}(\boldsymbol{X} + \boldsymbol{Z}_{R})^{-1}\boldsymbol{F}_{R}\left(\boldsymbol{H}\boldsymbol{v}_{G} + \boldsymbol{v}_{noise}\right) \tag{2.49}$$

where

$$\boldsymbol{H} = \boldsymbol{Z}_{RT}\boldsymbol{Z}_{TT} \tag{2.50}$$

$$\boldsymbol{v}_{noise} = \boldsymbol{v}_{SN} + \boldsymbol{F}_{R}^{-1}\left(\boldsymbol{Z}_{R}\boldsymbol{i}_{N} - \boldsymbol{v}_{N}\right) \tag{2.51}$$

with the component matrices [12, eq. (17)-(20)]

$$\boldsymbol{C}_{L} = \boldsymbol{Z}_{L}(\boldsymbol{Z}_{L} + \boldsymbol{Z}_{22amp})^{-1} \tag{2.52}$$

$$\boldsymbol{F}_{R} = \boldsymbol{Z}_{MR21}(\boldsymbol{Z}_{MR11} + \boldsymbol{Z}_{AR})^{-1} \tag{2.53}$$

$$\boldsymbol{Z}_{R} = \boldsymbol{Z}_{MR22} - \boldsymbol{F}_{R}\boldsymbol{Z}_{MR12} \tag{2.54}$$

$$\boldsymbol{X} = \boldsymbol{Z}_{11amp} - \boldsymbol{Z}_{12amp}(\boldsymbol{Z}_{22amp} + \boldsymbol{Z}_{L})^{-1}\boldsymbol{Z}_{21amp} \tag{2.55}$$

$$\boldsymbol{Z}_{TT} = \boldsymbol{F}_{T}^{\mathrm{T}}\left(\boldsymbol{Z}_{T} + \boldsymbol{Z}_{G}\right)^{-1} \tag{2.56}$$

$$\boldsymbol{F}_T = \boldsymbol{Z}_{MT12}(\boldsymbol{Z}_{MT22} + \boldsymbol{Z}_{AT})^{-1} \tag{2.57}$$

$$\boldsymbol{Z}_T = \boldsymbol{Z}_{MT11} - \boldsymbol{F}_T\boldsymbol{Z}_{MT21}. \tag{2.58}$$

We assume that $\boldsymbol{F}_R$ in (2.53) is invertible. Observe from (2.49) that $\boldsymbol{C}_L(\boldsymbol{X}+\boldsymbol{Z}_R)^{-1}\boldsymbol{F}_R$ multiplies both the signal and noise and is invertible. We thus focus on

$$\hat{\boldsymbol{v}}_L = \boldsymbol{H}\boldsymbol{v}_G + \boldsymbol{v}_{noise}. \tag{2.59}$$

We restrict attention to passive, lossless, and reciprocal matching networks, i.e., $\boldsymbol{Z}_{MR}$ has imaginary entries and $\boldsymbol{Z}_{MR} = -\boldsymbol{Z}_{MR}^{\mathrm{H}}$. The matching network impedance matrix thus has the form

$$\boldsymbol{Z}_{MR} = \begin{bmatrix} \boldsymbol{Z}_{MR11} & \boldsymbol{Z}_{MR12} \\ \boldsymbol{Z}_{MR12}^{\mathrm{T}} & \boldsymbol{Z}_{MR22} \end{bmatrix} = j \begin{bmatrix} \boldsymbol{X}_{MR11} & \boldsymbol{X}_{MR12} \\ \boldsymbol{X}_{MR12}^{\mathrm{T}} & \boldsymbol{X}_{MR22} \end{bmatrix} \tag{2.60}$$

where the $\boldsymbol{X}_{MRab}$ are real matrices, and $\boldsymbol{X}_{MR11}$ and $\boldsymbol{X}_{MR22}$ are symmetric.

2.4.1. Amplifier Noise

The amplifier voltage and current noise noise sources are modeled by Gaussian random variables with zero mean and second order statistics given by (see [12, eq. (10)])

$$\mathrm{E}\left[\boldsymbol{I}_N\boldsymbol{I}_N^{\mathrm{H}}\right] = \beta\boldsymbol{I} \tag{2.61}$$

$$\mathrm{E}\left[\boldsymbol{V}_N\boldsymbol{V}_N^{\mathrm{H}}\right] = \beta R_N^2\boldsymbol{I} \tag{2.62}$$

$$\mathrm{E}\left[\boldsymbol{V}_N\boldsymbol{I}_N^{\mathrm{H}}\right] = \rho\beta R_N\boldsymbol{I}. \tag{2.63}$$

Diagonal noise covariance matrices are reasonable if the amplifiers are well isolated on a chip, see [12], [4], [20] .

2.4.2. Antenna Noise

Suppose the background noise is caused by randomly polarized planar waves propagating from all angles uniformly. The open circuit noise voltage covariance then takes

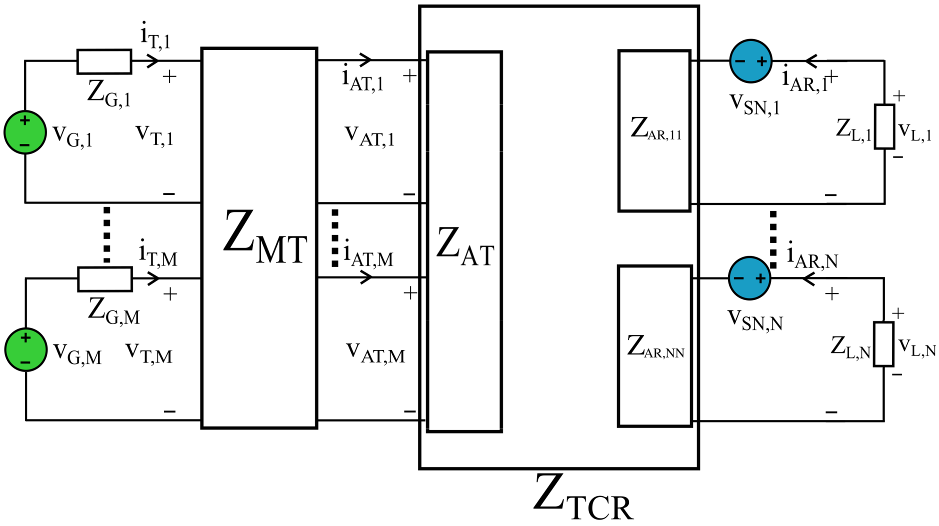


Figure 2.7.: Simplified MIMO front-end transceiver without coupling at the receiver

the form (see [12])

$$\mathrm{E}\left[\boldsymbol{V}_{SN}\boldsymbol{V}_{SN}^{\mathrm{H}}\right] = 4kT_{ant}W\Re\{\boldsymbol{Z}_{AR}\} \tag{2.64}$$

where $\Re\{\boldsymbol{Z}_{AR}\}$ is positive semi-definite. If there are additional losses then the antenna impedance is augmented by a real positive definite matrix $\boldsymbol{R}_{AL}$ at an equivalent noise temperature $T_{ant,L}$ to obtain

$$\mathrm{E}\left[\boldsymbol{V}_{SN}\boldsymbol{V}_{SN}^{\mathrm{H}}\right] = 4kT_{ant}W\Re\{\boldsymbol{Z}_{AR}\} + 4kT_{ant,L}W\boldsymbol{R}_{AL}. \tag{2.65}$$

To illustrate the effects of the correlations and matching, consider a model with only mutual transmitter coupling and no mutual coupling or antenna correlation at the receiver, see Fig. 2.7. We model the power amplifier outputs as generators with internal impedance and examine the effects of multiport matching on the capacity of a single user MIMO system with channel correlation. The results for two different matching approaches are shown in Fig. 2.8. One solution for decoupling and matching is given in closed form in (2.77), and the second, denoted in the figure as individual power matching, is the result of a global optimization over elements of a matching network that only connect individual antenna elements with their corresponding LNA. The

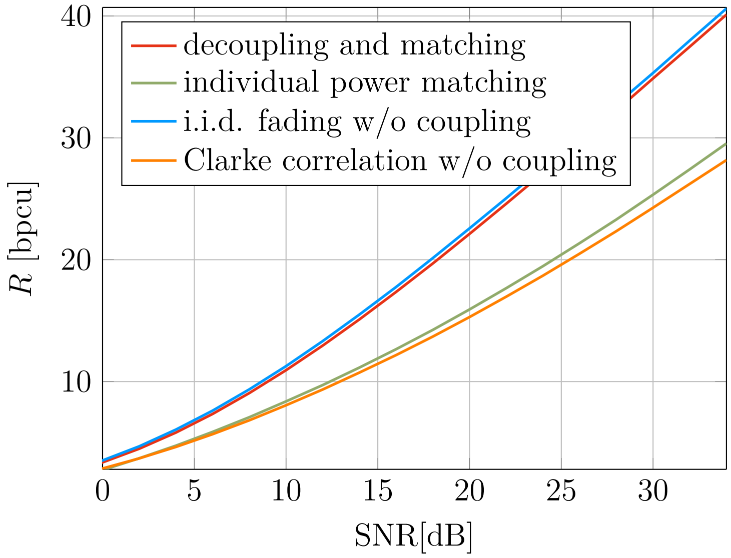


Figure 2.8.: Rates vs. SNR with receiver CSI

Clarke model correlation coefficients $\rho_{t/r}$ for the Kronecker spatial correlation matrix $\boldsymbol{R}_{t/r}$ are equal to $\mathrm{sinc}(\frac{2d}{\lambda})$, where d is the distance between two consecutive antenna elements and λ is the wavelength [22]. Interestingly, this simple correlation model captures very well the behaviour of the simulated antenna system with individual two-port antenna matching.

2.4.3. Capacity

Let the transmitter and receiver symbol samples be represented by $\boldsymbol{v}_G$ and $\hat{\boldsymbol{v}}_L$, respectively. Combining (2.59) and (2.50) we have

$$\hat{\boldsymbol{v}}_L = \boldsymbol{Z}_{RT}\boldsymbol{Z}_{TT}\boldsymbol{v}_G + \boldsymbol{v}_{noise}. \tag{2.66}$$

The mutual information of the source and load voltages with perfect receiver CSI is thus

$$\begin{aligned} I(\boldsymbol{V}_G; \hat{\boldsymbol{V}}_L) &= h(\hat{\boldsymbol{V}}_L) - \log_2\left((\pi e)^N \det(\boldsymbol{C}_{noise})\right) \\ &\leq \log_2 \det\left(\boldsymbol{I} + \boldsymbol{C}_{noise}^{-1}\boldsymbol{Z}_{RT}\boldsymbol{Z}_{TT}\boldsymbol{C}_{\boldsymbol{V}_G}\boldsymbol{Z}_{TT}^{\mathrm{H}}\boldsymbol{Z}_{RT}^{\mathrm{H}}\right) \end{aligned} \tag{2.67}$$

with equality if and only if $\boldsymbol{v}_G$ is a Gaussian distributed vector. Using (2.51) and (2.61)-(2.64), the noise covariance matrix is

$$\begin{aligned}\mathbf{C}_{noise} &= \mathrm{E}\left[\boldsymbol{V}_{SN}\boldsymbol{V}_{SN}^{\mathrm{H}}\right] + \boldsymbol{F}_R^{-1}\,\mathrm{E}\left[\left(\boldsymbol{Z}_R\boldsymbol{I}_N - \boldsymbol{V}_N\right)\left(\boldsymbol{Z}_R\boldsymbol{I}_N - \boldsymbol{V}_N\right)^{\mathrm{H}}\right]\mathbf{F}_R^{-\mathrm{H}} \\ &= 4kT_{ant}W\Re\{\mathbf{Z}_{\mathbf{AR}}\} + \beta\,\mathbf{F}_R^{-1}\left(\mathbf{Z}_R\mathbf{Z}_R^{\mathrm{H}} + R_N^2\boldsymbol{I} - 2R_N\Re\{\rho\mathbf{Z}_R^{\mathrm{H}}\}\right)\mathbf{F}_R^{-\mathrm{H}}. \qquad (2.68)\end{aligned}$$

The capacity is obtained by maximizing the mutual information over the transmitter and receiver matching networks, subject to the power constraints

$$\mathrm{Tr}(\boldsymbol{C}_{\boldsymbol{V}_G}) \leq P_{av} \qquad (2.69)$$

$$\mathrm{E}\left[\Re\{\boldsymbol{V}_G^{\mathrm{H}}\boldsymbol{C}_T\boldsymbol{V}_G\}\right] \leq P_{rad}. \qquad (2.70)$$

The constraint (2.69) limits the total supplied power which for decoupled antennas with perfect matching also constrains the radiated power. However, in a coupled MIMO system the supplied power is not necessarily the same as the radiated power, which is the quantity constrained by regulatory bodies. This has been highlighted in [20] that introduced the radiated power constraint (2.70).

To determine the optimal receiver matching network we use the following lemma.

Lemma 2.1. For a fixed $\boldsymbol{M} \succeq \mathbf{0}$ and $\boldsymbol{C}_1 \succeq \boldsymbol{C}_2 \succ \mathbf{0}$ we have

$$\log_2\det\left(\boldsymbol{I} + \boldsymbol{C}_2^{-1}\boldsymbol{M}\right) \geq \log_2\det\left(\boldsymbol{I} + \boldsymbol{C}_1^{-1}\boldsymbol{M}\right). \qquad (2.71)$$

Proof: According to [23, Theorem 7.2.6], there is a positive semidefinite matrix $\boldsymbol{M}^{1/2}$ such that $\boldsymbol{M}^{1/2}\boldsymbol{M}^{1/2} = \boldsymbol{M}$ and

$$\det\left(\boldsymbol{I} + \boldsymbol{C}_i^{-1}\boldsymbol{M}\right) = \det\left(\boldsymbol{I} + \boldsymbol{M}^{1/2}\boldsymbol{C}_i^{-1}\boldsymbol{M}^{1/2}\right) \qquad (2.72)$$

for $= 1, 2$. Moreover, by [23, Corollary 7.7.4] we have $\boldsymbol{C}_2^{-1} \succeq \boldsymbol{C}_1^{-1}$ and thus

$$\boldsymbol{I} + \boldsymbol{M}^{1/2}\boldsymbol{C}_2^{-1}\boldsymbol{M}^{1/2} \succeq \boldsymbol{I} + \boldsymbol{M}^{1/2}\boldsymbol{C}_1^{-1}\boldsymbol{M}^{1/2}.$$

The proof is completed by using [23, Corollary 7.7.4 (b)] which states that $\boldsymbol{A} \succeq \boldsymbol{B} \succ \mathbf{0}$ implies $\det\boldsymbol{A} \geq \det\boldsymbol{B}$. ■

By (2.67) and Lemma 2.1, we wish to find a smallest $\mathbf{C}_{noise}$ in the positive definite ordering. Using similar steps as in [4, Sec. V], one may rewrite (2.68) as

$$\begin{aligned}\mathbf{C}_{noise} &= 4kT_{ant}W\Re\{\mathbf{Z_{AR}}\} + \mathbf{F}_R^{-1}\beta\left((\mathbf{Z}_R - Z_{opt}\boldsymbol{I})(\mathbf{Z}_R - Z_{opt}\boldsymbol{I})^{\mathrm{H}}\right.\\ &\qquad \left.-2R_N\Re\{\rho\mathbf{Z}_R^{\mathrm{H}}\} + 2\Re\{Z_{opt}\mathbf{Z}_R^{\mathrm{H}}\}\right)\mathbf{F}_R^{-\mathrm{H}}\\ &= 4kT_{ant}W\Re\{\mathbf{Z_{AR}}\} + \mathbf{F}_R^{-1}\beta(\mathbf{Z}_R - Z_{opt}\boldsymbol{I})(\mathbf{Z}_R - Z_{opt}\boldsymbol{I})^{\mathrm{H}}\mathbf{F}_R^{-\mathrm{H}}\\ &\qquad + 4kT_0W(F_{min}-1)\Re\{\boldsymbol{Z}_{AR}\}\end{aligned} \tag{2.73}$$

where the last step follows by using (2.18), (2.19), and the lossless property of the matching network, i.e., the power at the input and output of the matching network is conserved with

$$\Re\{\boldsymbol{Z}_R\} = \mathbf{F}_R\Re\{\boldsymbol{Z}_{AR}\}\mathbf{F}_R^{\mathrm{H}}. \tag{2.74}$$

We thus have the following theorem.

Theorem 2.2. The capacity of the MIMO antenna system under consideration is achieved by a lossless, passive, and reciprocal matching network satisfying $\boldsymbol{Z}_R = Z_{opt}\boldsymbol{I}$ so that

$$C = \log_2 \det\left(\boldsymbol{I} + (\boldsymbol{C}_{noise}^{*})^{-1}\boldsymbol{Z}_{RT}\boldsymbol{Z}_{TT}\boldsymbol{C}_{\mathbf{V}_G}\boldsymbol{Z}_{TT}^{\mathrm{H}}\boldsymbol{Z}_{RT}^{\mathrm{H}}\right) \tag{2.75}$$

where

$$\boldsymbol{C}_{noise}^{*} = 4kT_0W\Re\{\mathbf{Z_{AR}}\}\left(F_{min} - 1 + \frac{T_{ant}}{T_0}\right). \tag{2.76}$$

Proof: From (2.73) we have $\boldsymbol{C}_{noise} \succeq \boldsymbol{C}_{noise}^{*}$ with equality if $\boldsymbol{Z}_R = Z_{opt}\boldsymbol{I}$, i.e., the array has been effectively decoupled. Now apply Lemma 2.1 to (2.67). There is, in general, a *class* of lossless, passive, and reciprocal $\mathbf{Z}_{MR}$ that achieve capacity. For instance, one may choose (see [4, 12] and also [24])

$$\mathbf{Z}_{MR} = j\begin{bmatrix} -\Im\{\boldsymbol{Z}_{AR}\} & (\Re\{\mathbf{Z}_{AR}\}\Re\{Z_{opt}\})^{1/2} \\ (\Re\{\boldsymbol{Z}_{AR}\}\Re\{Z_{opt}\})^{1/2} & \Im\{Z_{opt}\}\boldsymbol{I}\end{bmatrix}. \tag{2.77}$$

Inserting $\mathbf{Z}_{MR}$ into (2.54) with (2.53) we have the desired decoupling. ■

Theorem 2.2 effectively appears in [4] and Lemma 2.1 provides an alternative and slightly simpler proof.

2.5. Sensitivity Analysis

We next evaluate the sensitivity of the above matching circuits by varying the device tolerances, the bandwidth, and the antenna spacing.

Antennas

Suppose both the transmit and receive arrays are uniform linear arrays (ULA) with half-wavelength (resonant) dipoles with center feed oriented in parallel to each other. Closed form expressions exist for the self and mutual impedance of very thin wire dipoles [15], however no such expressions exist for radiation patterns. This motivates evaluating the antenna array impedance matrix and patterns using a numerical method of moments (MoM) provided by the Antenna Toolbox in Matlab and benchmarked against 4nec2 [25] software. We use dipoles of length $\lambda/2$ and width $\lambda/100$ separated by spacings no smaller than 0.05λ. We evaluate the antenna properties at the center frequency $f_c = 800$MHz.

Noise Parameters

Consider amplifiers with $R_N = 10\,\Omega$, $Z_{opt} = 56.74 + j10.66$, and minimum noise figure $F = 1.36\,$dB. These parameters are motivated by perfectly unilateral amplifiers with $Z_{amp12} = 0\,\Omega$, $|Z_{amp21}| >> 1$, and $Z_{amp11} = Z_{opt}^*$, i.e., we consider

$$Z_{amp} = \begin{bmatrix} Z_{opt}^* & 0 \\ Z_{amp21} & Z_{amp22} \end{bmatrix}. \tag{2.78}$$

Such models do not depart much from well-designed catalog amplifiers used in [4, 20], for example. Note, however, that Z_{amp} does not affect the capacity calculation.

Information Rates

Consider CSI at the receiver while the transmitter knows only the statistics of the channel. In this case isotropic diagonal power allocation $\boldsymbol{C}_{\boldsymbol{V}_G} = \frac{P}{N}\boldsymbol{I}$, where P is the transmitter power and N is the number of transmit/receive antennas, is the best strategy [18].

For each antenna spacing we evaluate the rate by Monte Carlo simulation with 25,000 channel realizations. We compare the optimal receive matching rates with:

- ▷ independent and identically distributed (iid) fading and noise; the fading is Rayleigh and flat, and the receiver noise is spatially white;
- ▷ self-matching, i.e., the dipole antennas are matched to the optimal noise impedance of the amplifier so that

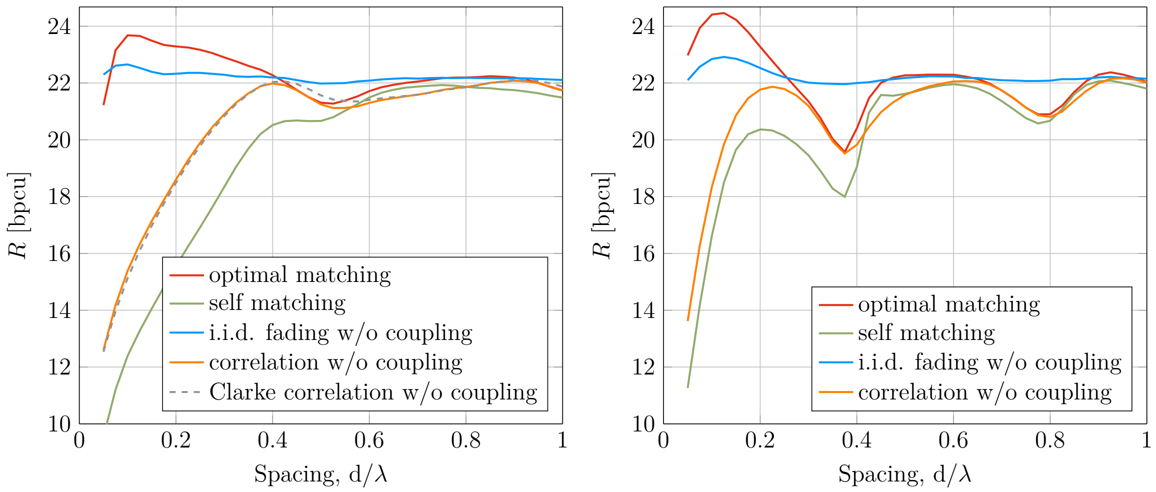
$$\mathbf{Z}_{MR,\mathrm{self},ab} = j\,\mathrm{diag}\left(\mathbf{X}_{MRab}\right) \tag{2.79}$$

where diag$(\cdot)$ retains the diagonal of a matrix.

For fair comparison we scale the channel matrices so that $\mathrm{E}\left[\mathrm{Tr}(\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}})\right] = N^2$. The SNR is defined with respect to the SISO SNR. For uncoupled and uncorrelated MIMO RF chains the individual chain SNR is equal to our definition of the SISO SNR.

2.5.1. Antenna Spacing

Consider $N = 4$ dipole antenna elements. Fig. 2.9 shows the rate as a function of the spacing between antennas at the receive array for the cases described above. The SNR is fixed at 20dB. Both uniform linear and circular arrays are considered. For small antenna spacings the rates exceed those of i.i.d. fading and noise because of the larger power collected by the decoupled and matched array through its larger effective aperture as compared to the uncoupled array (e.g., see [26]). This is reflected also in our simulations by having the SNR benchmarked to the isolated dipole impedance, whereas the coupled dipoles' self impedance can be considerably larger. We further



(a) ULA

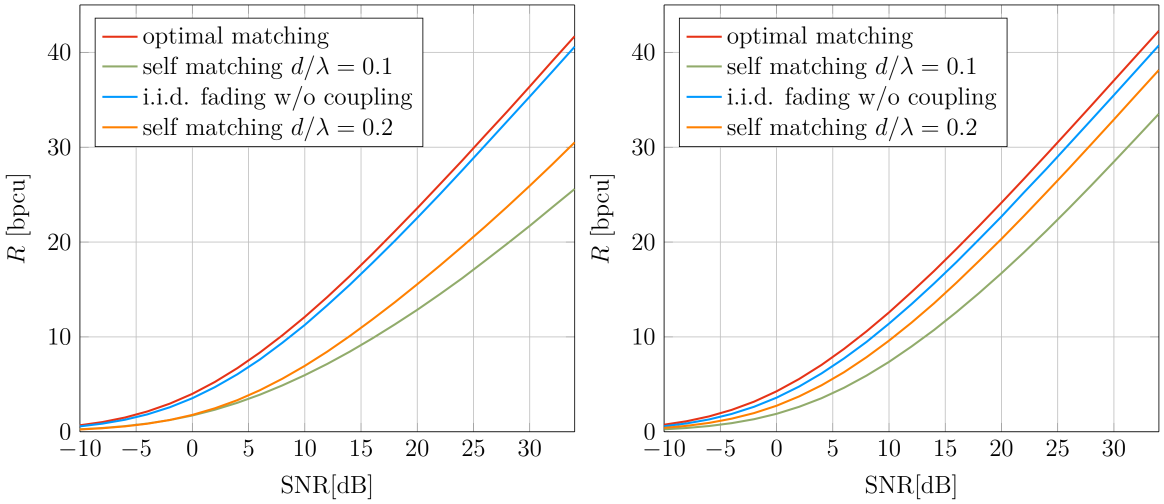
(b) UCA

Figure 2.9.: Rate vs. antenna spacings for various matching strategies

observe that the results are qualitatively consistent for both array architectures, hinting that the placement of the dipoles in a 2-D plane does not fundamentally change the results, e.g., the ordering of the rates at a given spacing. Furthermore, for the uniform circular array (UCA) and a specified information rate, the antenna spacing is smaller than for the ULA. We expect that optimizing the positions of the antenna elements will improve these results.

Fig. 2.10 plots the information rate vs. SNR for $d = 0.1\lambda$ and $d = 0.2\lambda$ for both UCAs and ULAs. At low SNR the gap between suboptimal and optimal matching is not significant and therefore closer spacing can be used with suboptimal matching without rate penalties. However, at high SNR the gap is significant. For example, for UCA at a spacing of 0.1λ and a rate of 15 bits/s/Hz there is a SNR loss of 7 dB as compared to optimal matching. Again, the slope of the information rate vs SNR is steeper for UCA than for ULA for the suboptimal self-matching strategy. This shows that both antenna-matching network co-design and optimizing the positions of the antenna elements is of practical relevance.

We remark that an optimal decoupling network is complex and may result in a large and bulky front-end. An optimal dense matching network has a complexity of $2N^2 + N$ network elements and requires connections between all pairs of antennas.



(a) ULA

(b) UCA

Figure 2.10.: Rate vs. SNR for various matching strategies

For lossless and reciprocal networks, which we focus on in this thesis, the number of components can be reduced by N^2. Research into low complexity implementations of such networks is presented in [24] and references therein. However, there are still strong limitations in bridging the gap to practicality for such structures and medium and large size arrays. This motivates a joint antenna and matching co-design as a viable complexity performance trade-off.

2.5.2. Device Variations

The receiver matching network was derived by maximizing the mutual information at one frequency. However, most applications require operating over a large spectral range. In addition, realistic components will cause the entries of the matching network to differ from the desired ones because of losses, parasitic effects, availability of only a discrete set of nominal values (e.g. for lumped elements), fabrication tolerances, temperature and aging effects. We investigate the robustness of the matching network to device variations.

We evaluate the rates for a specific matching circuit architecture, namely the generalized Π network shown in Fig. 2.11. This network is taken as a benchmark because

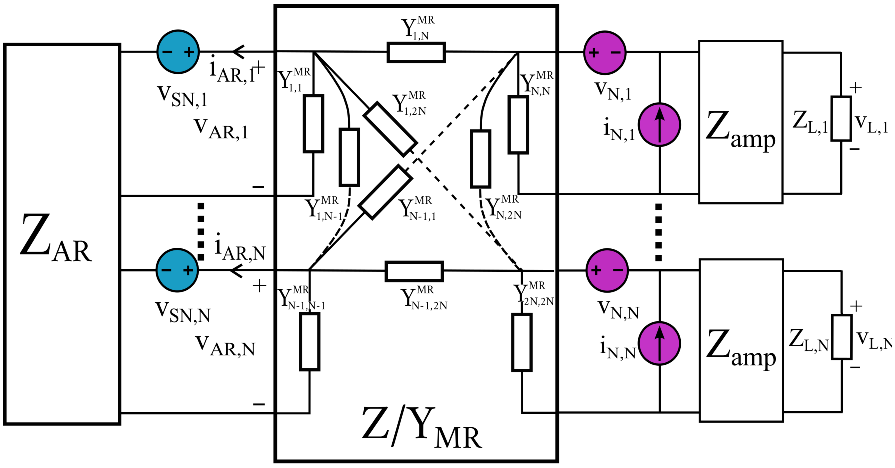


Figure 2.11.: MIMO receiver with a generalized Π matching network

it naturally extends classical Π networks from SISO matching as in Fig. 2.12. The two-port Π network admittance matrix is given by:

$$\boldsymbol{Y}_{\Pi} = \begin{bmatrix} Y_{1,1} + Y_{1,2} & -Y_{1,2} \\ -Y_{1,2} & Y_{2,2} + Y_{1,2} \end{bmatrix}. \tag{2.80}$$

Similarly, the generalized $2N$-port Π network has an admittance matrix:

$$\boldsymbol{Y}_{gen-\Pi} = \begin{bmatrix} \sum_{j=1}^{2N} Y_{1,j} & -Y_{1,2} & \cdots & -Y_{1,2N} \\ -Y_{1,2} & \sum_{j=1}^{2N} Y_{2,j} & \cdots & -Y_{2,2N} \\ \vdots & \vdots & \vdots & \vdots \\ -Y_{1,2N} & -Y_{2,2N} & \cdots & -Y_{2N,2N} \end{bmatrix}. \tag{2.81}$$

If we consider purely reactive networks we only have $Y_{i,i} = \frac{1}{2\pi f_c L_{i,i}}$ or $Y_{i,i} = 2\pi f_c C_{i,i}$ for inductors and capacitors at the design frequency f_c. We thus perturb only the imaginary part of our lumped network components, but in practice we may also have resistive parasitics.

Fig. 2.13 plots the average information rate as a function of antenna spacing at an SNR of 20dB. The tolerances have been chosen to be equal for all components, both capacitors or inductors. The value of individual components with tolerance is computed as $Y_{L,C} = \mathrm{Y}_{\mathrm{L,C}_{\mathrm{n}}\mathrm{ominal}}(1 + x * \mathrm{tol.})$ where $x \sim \mathcal{N}(0, 1)$. We use only one channel

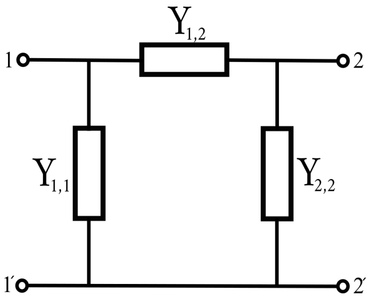
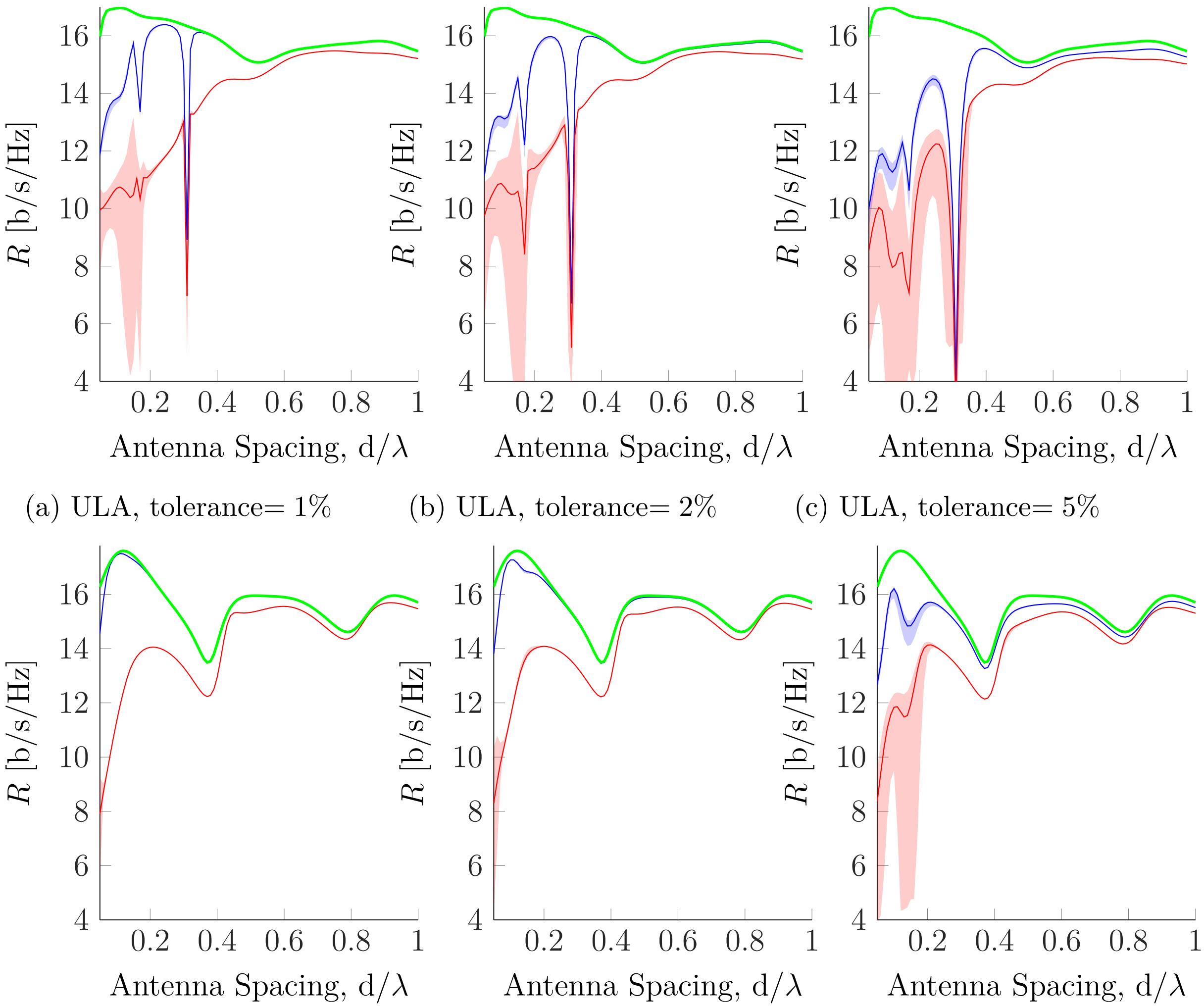


Figure 2.12.: Two-port Π network with lumped elements

realization that is chosen at random and then reused for all our experiments. We generate 50,000 random samples for each perturbed component to obtain statistically significant results.

Fig. 2.13 plots information rates for the unperturbed optimal matching network, and the average and range of information rates evaluated by sampling from the elements with component tolerances. We also perturb the self matching solution for comparison. The solid curves are the empirical means of the collected statistics. The shaded areas are generated by collecting the realizations into two classes: those that fall above and those that fall below the empirical mean. We then compute the variances of the samples of these two classes. The shaded area shows the values between the mean plus the variance of the "above the mean" curves, and the mean minus the variance of the "below the mean" curves. The resulting region is thus asymmetric around the mean value in general and the plot emphasizes the heavier tails one one side of the distribution.

We observe that the UCA is insensitive to perturbations up to tolerance values of 5%, while the ULA shows significant performance degradation even at 1%. While perturbations only decrease rates for optimum matching, self-matching can benefit with higher rates than for the unperturbed case. This is expected, since self matching is a sub-optimal diagonal matching. Optimal networks exhibit lower rates with close antenna spacing with a given level of perturbation, but the variations are consistently low for both ULA and UCA, which is not the case for self matching. This seems to be an unexpected result of practical importance, as self matching appears more sensitive to perturbations (larger variance). For antenna spacings larger than $d/\lambda \geq 0.5$ the



(a) ULA, tolerance= 1%

(b) ULA, tolerance= 2%

(c) ULA, tolerance= 5%

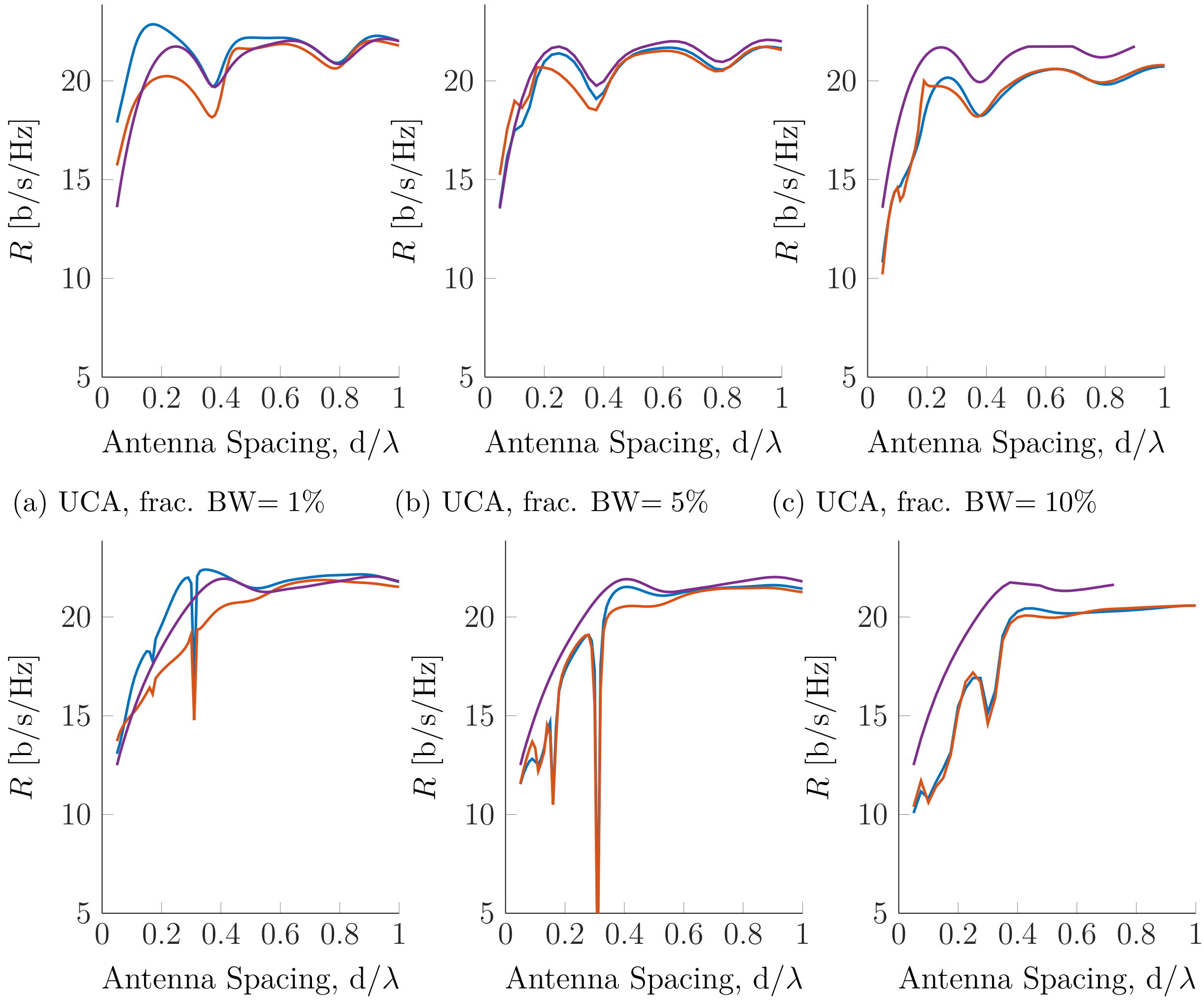
(d) UCA, tolerance= 1%

(e) UCA, tolerance= 2%

(f) UCA, tolerance= 5%

Figure 2.13.: UCA and ULA information rates with element tolerances. The unperturbed optimal rates are given in solid green. Average perturbed optimal matching (blue solid line), range for perturbed optimal matching (blue shaded area); average perturbed self matching (red solid line), range for perturbed self matching (red shaded area).

effects of perturbations are much less significant for both matching strategies, as the effects of mutual coupling are reduced.



(a) UCA, frac. BW= 1%

(b) UCA, frac. BW= 5%

(c) UCA, frac. BW= 10%

(d) ULA, frac. BW= 1%

(e) ULA, frac. BW= 5%

(f) ULA, frac. BW= 10%

Figure 2.14.: Broadband information rates with the optimal matching network designed at the central frequency (blue solid line); with the self-matching network designed at the central frequency (red solid line); and with channel correlations but without antenna coupling (purple solid line).

2.5.3. Broadband Rates

Fig. 2.14 shows information rates as a function of antenna spacing at two-sided factional bandwidths of 1%, 5% and 10% of the carrier frequency f_c. The receiver optimal matching network is computed for the parameters found at the central frequency. The bandwidth is divided in $K = 200$ equally spaced bands, and the mid-frequency in each band is denoted as f_k. Therefore the total rate per unit of bandwidth is

$$I(\boldsymbol{V}_G; \hat{\boldsymbol{V}}_L) = \frac{1}{K} \sum_{k=1}^{K} \log_2 \det \left(\boldsymbol{I} + \boldsymbol{C}_{noise,f_k}^{-1} \boldsymbol{Z}_{RT,f_k} \boldsymbol{Z}_{TT,f_k} \boldsymbol{C}_{\boldsymbol{V}_G} \boldsymbol{Z}_{TT,f_k}^{\mathrm{H}} \boldsymbol{Z}_{RT,f_k}^{\mathrm{H}} \right) \tag{2.82}$$

where all other channel and network parameters except the matching network are evaluated at the equispaced frequencies f_k over which the summation is done.

Fig. 2.14 emphasizes that the optimal matching network is highly frequency selective. This is expected since the matching network is a function of the antenna array impedance matrix that changes with frequency. The optimal matching networks derived in the previous chapter do not account for the frequency dependence. These aspects are outside the scope of this thesis but are studied in [5, 27–30] for example.

2.6. Conclusions

In this chapter we have shown that the information-rate optimal design of receivers with lossless, passive, and reciprocal matching networks for MIMO RF front-ends can be solved in closed form. The optimal networks form a parametrized family of circuits. Our models consider both signal and noise correlations as well as all circuit theoretic interactions relevant for a compact antenna array with mutual coupling. The results shows that noise covariance minimization in the positive definite matrix sense is equivalent to mutual information maximization with Gaussian noise and Gaussian signaling. Interestingly, this is a vector extension of the single antenna case where mutual information maximization is equivalent to minimization of the noise figure of the receiver. This is true irrespective of the optimization of the signal covariance at the transmitter, since there are no non-positive eigenvalues of the noise covariance of a physical system (all signal directions contain some noise, and the optimization decreases noise in all directions). An optimally matched receiver is one where the coupled array is simultaneously decoupled, and where each LNA is individually presented with the minimum noise input impedance at the input port. Both signal and noise paths are decoupled and the resulting front end can be seen as a collection of independent RF chains, one for each antenna element.

The performance of the matching networks is, however, sensitive to device manufacturing variations. For example, depending of the architecture of the matching network and the design of the antenna array, a tolerance of 5% in each lumped element in the matching network can degrade the rate by 10% to 30% at an antenna spacing of 0.2λ. It is also noted that antenna array design can significantly influence the rates. For instance, the antennas can be spaced more closely for a UCA than for a ULA for a specified information rate.

Finally, the matching networks are no longer optimal if we consider transmission in larger bandwidths. A system operating in a fractional bandwidth of 10% around the carrier frequency does not benefit from optimal matching and supports only the same information rates as single port self matching for both ULA and UCA arrays.

3

Precoding for Multi-User MIMO with Discrete Signaling

3.1. Introduction

Massive multiple input multiple output (massive MIMO) uses large antenna arrays at base stations to serve many users that each have a small number of antennas [31]. The gains of massive MIMO include improved power and spectral efficiencies, and simplified signal processing [32]. The gains are often stated for a large number N of base station antennas and a large number K of User Equipment (UE) when the ratio N/K is held constant.

An implementation for large N and K is challenging. For example, consider a base station deployment with N radio frequency (RF) chains. It seems impractical from a cost and power consumption point of view to use high-resolution analog-to-digital and digital-to-analog converters (ADCs/DACs) along with linear but low-efficiency power amplifiers. In fact, the papers [33] and [34] argue that high resolution DACs and linear power amplifiers not only dominate the power consumption of Massive MIMO base-stations (up to 75% of the total power) but also account for major heat dissipation bottlenecks.

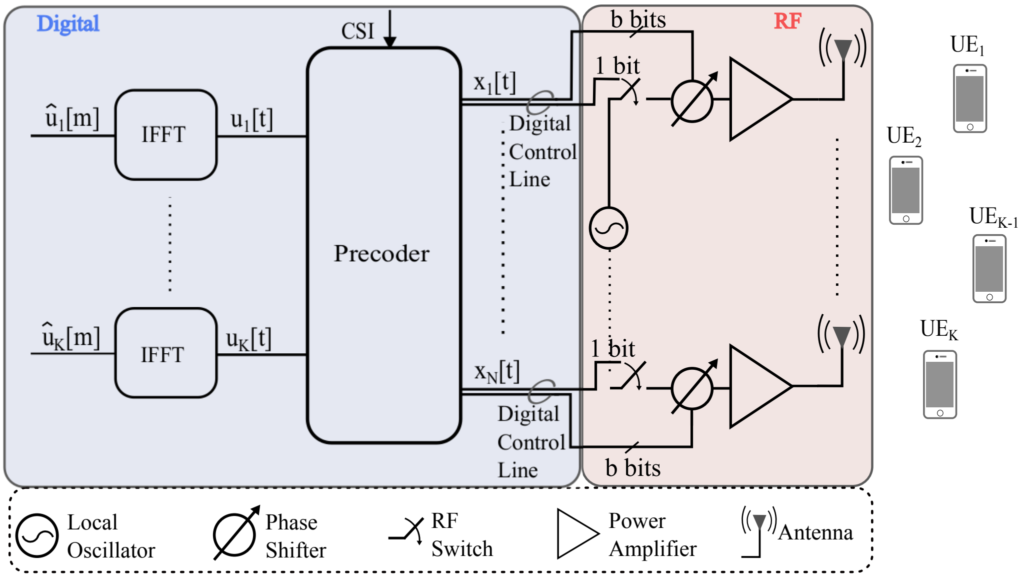


Figure 3.1.: Multi-user MIMO downlink with a low resolution digitally controlled analog architecture.

There are two main directions of research aimed at obtaining practical solutions for implementing massive MIMO. First, hybrid-beamforming [35] uses analog beamformers in the RF chain of each antenna, and the digital baseband processing is shared among different RF chains. This solution is mainly targeted at use cases in very high frequency bands (millimeter wave). Second, low-resolution ADCs/DACs or low resolution digitally controlled RF chains simplify the transceivers, e.g., one bit quantizers use simple comparators and they permit using non-linear power amplifiers combined with quasi-constant envelope waveforms.

We follow the second approach and propose a transceiver architecture and nonlinear baseband precoding algorithms to obtain a low-power, low-cost and high efficiency RF co-design. We propose to eliminate the DACs and replace them with low resolution digitally controlled phase-shifters that can be used together with high efficiency nonlinear amplifiers [36].

The high-level joint digital-analog architecture is shown in Fig. 3.1. The Inverse fast Fourier Transform (IFFT) and the Precoder are implemented fully in the digital domain, while the Local Oscillator (LO), programmable phase shifter, RF switch and Power Amplifier (PA) are RF components. The phase shifter is actuated directly from the digital domain with b bits without mixed signal components such as DACs, and operates on the signal generated by a local oscillator. The phase shifter is placed before a power amplifier to both mitigate its insertion loss and to simplify the matching of the PA and the antenna. We emphasize that the phase shifter is in general operated at baseband sampling rate, thus considerably reducing the switching time. The RF switch turns individual chains on and off as specified by the precoder in order to mitigate self-interference, this will be reflected in the choice of signaling alphabet.

3.2. System Model

Fig. 3.1 depicts the downlink of a Multi-User MIMO (MU-MIMO) channel with N transmit antennas and K UEs that each have a single antenna. A discrete-time, frequency selective, time-varying, baseband channel has a finite impulse response filter between each pair of transmit and receive antennas. We collect the received signals $y_k[t]$ at time t, $t = 1, 2, \ldots, T$, of user k, $k = 1, 2, \ldots, K$, into the K-dimensional column vector $\boldsymbol{y}[t] = [y_1[t], \ldots, y_K[t]]^{\mathrm{T}}$. We have

$$\boldsymbol{y}[t] = \sum_{l=0}^{L-1} \boldsymbol{H}[t, l]\boldsymbol{x}[t-l] + \boldsymbol{z}[t] \tag{3.1}$$

where $\boldsymbol{x}[t] = [x_1[t], \ldots, x_N[t]]^{\mathrm{T}}$ is the N-dimensional transmit vector, the $\boldsymbol{H}[t, l]$, $\ell = 0, 1, \ldots, L-1$, are time-varying $K \times N$ channel matrices, and $\boldsymbol{z}[t]$ is a circularly-symmetric, complex, Gaussian, column vector with a scaled identity covariance matrix, i.e., we have $\boldsymbol{z} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\boldsymbol{I})$. We write $\boldsymbol{x}$ for $\boldsymbol{x}[t]$ when the time parameter is clear from the context. Power constraints are given by one or more of the following inequalities: an average power constraint $\mathrm{E}\left[\|\boldsymbol{x}\|^2\right] \leq P$, an instantaneous power constraint $\|\boldsymbol{x}\|^2 \leq P$ with probability one, or individual power constraints $|x_n|^2 \leq P/N$, $n = 1, 2, \ldots, N$, with probability one. These constraints often allow closed form expressions to optimization problems, and efficient numerical algorithms that can find solutions that are globally optimal.

Equation (3.1) represents the Shannon-Nyquist sampled input-output relationship of a time-varying multi-path communication channel. For example, a single tap channel has $L = 1$ and with time-variations we recover the flat fading channel model

$$\boldsymbol{y}[t] = \boldsymbol{H}[t]\boldsymbol{x}[t] + \boldsymbol{z}[t]. \tag{3.2}$$

Another interesting model is the block fading channel where the channel matrix changes every T channel uses:

$$\begin{aligned}\boldsymbol{y}[kT+1:(k+1)T] = \boldsymbol{H}[kT+1]\,\boldsymbol{x}[kT+1:(k+1)T]\\ +\,\boldsymbol{z}[kT+1:(k+1)T]\end{aligned} \tag{3.3}$$

where k is an integer representing the k-th block, T is the block length, and we use the notation $\boldsymbol{x}[t:t+T-1] = [\boldsymbol{x}[t], \ldots, \boldsymbol{x}[t+T-1]]$.

For the remainder of this chapter we study block fading channels. We write

$$\boldsymbol{H}[t,l] = \boldsymbol{H}[l] = \begin{pmatrix} h_{11}[l] & h_{12}[l] & \ldots & h_{1N}[l] \\ h_{21}[l] & h_{22}[l] & \ldots & h_{2N}[l] \\ \vdots & \vdots & \ddots & \vdots \\ h_{K1}[l] & h_{K2}[l] & \ldots & h_{KN}[l] \end{pmatrix} \tag{3.4}$$

for a generic block, where $h_{kn}[l]$, $l = 0, 1, \ldots, L-1$, is the channel impulse response from the n-th antenna at the base station to the k-th UE. We consider a Rayleigh fading frequency selective channel with a uniform power delay profile, i.e., we choose $\mathrm{E}\left[|h_{kn}[l]|^2\right] = \frac{1}{L}$, where the $h_{kn}[l] \sim \mathcal{CN}(0, \frac{1}{L})$ are iid circularly-symmetric, complex Gaussian random variables. We mostly assume that the realizations $\boldsymbol{H}[l]$, $l = 0, 1, \ldots, L-1$, are known perfectly at the transmitter (CSIT) but we do study the effects of approximate knowledge in Sec. 3.8 below.

Finally, we give the discrete frequency domain formulation for the channel input-output relation since precoding for frequency selective channels is usually performed on a per-subcarrier basis. Let $\hat{\boldsymbol{x}}[m]$ and $\hat{\boldsymbol{y}}[m]$, $m = 1, \ldots, T$, denote the frequency-domain transmit and receive vectors in (3.1) when using OFDM with a cyclic prefix and a size T DFT. The new channel can be expressed as

$$\hat{\boldsymbol{y}}[m] = \hat{\boldsymbol{H}}[m]\hat{\boldsymbol{x}}[m] + \hat{\boldsymbol{z}}[m] \tag{3.5}$$

where $\hat{\boldsymbol{H}}[m]$ is a $K \times N$ matrix representing the frequency domain channel matrix per subcarrier:

$$\hat{\boldsymbol{H}}[m] = \sum_{l=0}^{L-1} \boldsymbol{H}[l] e^{-\mathrm{j}\, 2\pi l(m-1)/T}, \quad m = 1, \ldots, T \tag{3.6}$$

and where the noise $\hat{\boldsymbol{z}}[m]$ corresponds to the frequency domain transformation of $\boldsymbol{z}[t]$.

3.2.1. Modulation and Receiver Metric

We consider $\boldsymbol{x}$ with entries taken from a discrete alphabet $\mathcal{X}$ that has $2^b + 1$ elements where b bits encode the phase of each entry x_n. More precisely, we choose

$$\mathcal{X} = \{0\} \cup \left\{ \sqrt{\frac{P}{N}} e^{\mathrm{j}\, 2\pi q/2^b}; q = 0, 1, \ldots, 2^b - 1 \right\}. \tag{3.7}$$

We select $P = 1$ and define $\mathrm{SNR} = 1/\sigma^2$. The alphabet (3.7) permits per-symbol antenna selection through the $\{0\}$ symbol. The idea of joint precoding and antenna selection also appeared in [37] but our algorithms will select antennas without enforcing sparsity. The 0 symbol corresponds to a digitally controlled RF switch that can turn individual antennas on and off.

We use the mean squared error (MSE) as a precoding metric. For example, consider the fading model (3.2) with $\boldsymbol{H}[t] = \boldsymbol{H}$. The MSE between a target signal $\boldsymbol{u}$ and its estimate $\bar{\boldsymbol{u}} = \alpha \boldsymbol{y} = \alpha(\boldsymbol{H}\boldsymbol{x} + \boldsymbol{z})$ at the receiver is

$$MSE = \mathrm{E}\left[\|\boldsymbol{u} - \bar{\boldsymbol{u}}\|_2^2 \right] = \|\boldsymbol{u} - \alpha \boldsymbol{H}\boldsymbol{x}\|_2^2 + \alpha^2 K \sigma^2. \tag{3.8}$$

This choice is motivated by the massive MIMO literature [38] that shows that precoders designed with simple criteria can approach capacity. Also, the MSE is related to the mutual information [39] for Gaussian channels.

3.2.2. Flat Fading Channels

Let $u_k \in \mathcal{U}$ be the complex symbol that we wish to generate at the k-th UE, $k = 1, \ldots, K$, where $\mathcal{U}$ is either a 4-, 16-, or 64-Quadrature Amplitude Modulation (QAM) signaling set. Let $\boldsymbol{u}$ be a column vector with K symbols. Consider the precoding problem

$$\begin{aligned} \min_{\boldsymbol{x},\alpha} \quad & \|\boldsymbol{u} - \alpha \boldsymbol{H}\boldsymbol{x}\|_2^2 + \alpha^2 K \sigma^2 \\ \text{s.t.} \quad & \boldsymbol{x} \in \mathcal{X}^N \\ & \alpha > 0. \end{aligned} \tag{3.9}$$

The factor α permits trading off noise enhancement and the received signal power. For example, the latter is more important at low SNR when the MSE is minimized by the Matched Filter (MF) [40]. Problem (3.9) is not a convex optimization program, unlike the classical linear MMSE counterpart, where $\alpha = 1$ and $\mathcal{X} = \mathbb{C}$. In particular, the constraint set is a bounded discrete Cartesian product set, thus making our problem a combinatorial problem.

3.2.3. Block Fading

The transmitter must compute $\boldsymbol{x}$ and α for each transmit vector $\boldsymbol{u}$. The precoding factor α can be broadcast to the receiver through a control channel but such a broadcast channel is not necessarily available. Moreover, the number of distinct α values is generally large for a large number of users and for large modulation sets. We approach the problem by studying three scenarios.

- ▷ Scenario 1: The base station computes a pair $(\alpha, \boldsymbol{x})$ for each $\boldsymbol{u}$ and broadcasts all precoding factors to the receivers. This scenario seems practically unrealistic but provides a performance benchmark.
- ▷ Scenario 2: The base station computes a pair $(\alpha, \boldsymbol{x})$ for each $\boldsymbol{u}$ and the receiver estimates an auxiliary channel as described in Sec. 3.3.3 below. This scenario requires no side information on the precoding factors.
- ▷ Scenario 3: The base station computes one α per coherence interval and the receiver estimates an auxiliary channel as described in Sec. 3.3.3 below.

The precoding factor α turns out to be mostly relevant for hard detection where proper scaling is needed to apply symbol-wise decision thresholds. For soft-decision receivers that generate log likelihood ratios (LLRs) this factor impacts mainly decoders that are sensitive to LLR magnitude scaling. In [41], the authors also discuss block fading and different estimation strategies.

For the block fading channel in (3.3) the precoding problem with $\boldsymbol{H} = \boldsymbol{H}[1]$ is

$$\begin{aligned} \min_{\boldsymbol{x}[1],\ldots,\boldsymbol{x}[T],\alpha} \quad & \sum_{t=1}^{T} \|\boldsymbol{u}[t] - \alpha \boldsymbol{H}\boldsymbol{x}[t]\|_2^2 + \alpha^2 T K \sigma^2 \\ \text{s.t.} \quad & \boldsymbol{x}[t] \in \mathcal{X}^N,\ t = 1, 2, \ldots, T. \\ & \alpha > 0. \end{aligned} \tag{3.10}$$

3.2.4. Frequency Selective Channels

For time-invariant frequency selective channels (3.1) with $\boldsymbol{H}[t,l] = \boldsymbol{H}[l]$, the precoder outputs a string of column vectors $\boldsymbol{x}[1], \ldots, \boldsymbol{x}[T]$ where T is the OFDM symbol length. In practice, T is chosen to balance the need for accurate channel state information (CSI) and quality-of-service (QoS) requirements.

Consider the target vectors $\boldsymbol{u}[1], \ldots, \boldsymbol{u}[T]$. The optimization problem is now

$$\begin{aligned} \min_{\boldsymbol{x}[1],\ldots,\boldsymbol{x}[T],\alpha} \quad & \sum_{t=1}^{T} \left\| \boldsymbol{u}[t] - \alpha \sum_{l=0}^{L-1} \boldsymbol{H}[l]\boldsymbol{x}[t-l] \right\|_2^2 + \alpha^2 T K \sigma^2 \\ \text{s.t.} \quad & \boldsymbol{x}[t] \in \mathcal{X}^N,\ t = 1, \ldots, T \\ & \alpha > 0. \end{aligned} \tag{3.11}$$

The problem (3.11) suggests a time domain approach rather than the frequency domain approach of [42]. The main advantage of the former approach is that one does not need to switch between time and frequency domains to enforce the discrete alphabet constraint (3.7). The cost of each such shift is a length T discrete Fourier transform.

3.3. Multiuser MIMO

3.3.1. Broadcast Channel

MU-MIMO has a base station serving multiple terminals on the same time-frequency resources by exploiting spatial degrees of freedom. The idea dates back at least to [43] and the information theoretic limits of MU-MIMO have been examined in [44–46] and many other papers.

The baseband received signal at user k for the Gaussian MIMO Broadcast Channel (GMBC) with block fading is given by (3.2)

$$y_k[t] = \boldsymbol{h}_k^{\mathrm{T}} \boldsymbol{x}[t] + z_k[t], \quad t = 1, 2, \ldots, T. \tag{3.12}$$

where $\boldsymbol{h}_k^{\mathrm{T}} = [h_{k1}, \ldots, h_{kN}]$ is the channel vector corresponding to user k. The capacity region of the time-invariant GMBC is achieved by Gaussian input distributions with an optimized covariance matrix $\boldsymbol{Q}$. The sum-rate optimization problem simplifies to diagonal $\boldsymbol{Q}$ with entries defined by the power allocation vector $\boldsymbol{q} = [q_1, \ldots, q_N]^{\mathrm{T}}$. The best sum-rate is

$$\begin{aligned} R_{sum} =& \max_{\boldsymbol{q}} \quad \log_2 \det \left(\boldsymbol{I} + \boldsymbol{H}\boldsymbol{Q}\boldsymbol{H}^{\mathrm{H}} \right) \\ & \text{s.t.} \quad \mathrm{Tr}(\boldsymbol{Q}) \leq P_{total} \\ & \qquad\ \ \boldsymbol{Q} \succeq 0 \end{aligned} \tag{3.13}$$

where

$$\boldsymbol{Q} = \begin{bmatrix} q_1 & 0 & \cdots & 0 \\ 0 & q_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & q_N \end{bmatrix}.$$

However, for massive MIMO it usually suffices to consider uniform power allocation, and we obtain

$$R_{sum} = \log_2 \det \left(\boldsymbol{I} + \frac{P}{K} \boldsymbol{H}\boldsymbol{H}^{\mathrm{H}} \right) \leq C. \tag{3.14}$$

If coding is performed across many realizations of a stationary and ergodic channel, then the sum capacity with perfect CSI at the receiver is given by [47,48]:

$$\begin{aligned} &\mathrm{E}\left[\max_{\boldsymbol{Q}} \quad \log_2 \det\left(\boldsymbol{I} + \boldsymbol{H}\boldsymbol{Q}\boldsymbol{H}^{\mathrm{H}}\right)\right] \\ &\text{s.t.} \qquad \mathrm{Tr}(\boldsymbol{Q}) \leq P_{total} \\ &\qquad\qquad \boldsymbol{Q} \succeq 0. \end{aligned} \tag{3.15}$$

The optimization can be computed with standard convex optimization tools [49,50].

It was shown in [44] that the sum capacity in (3.13) can be achieved by Dirty Paper Coding (DPC). The main idea of DPC consists of coding against known interference at the transmitter [44,51,52]. Unfortunately, DPC currently seems to be too complex even for medium size systems. Also, the presence of perfect CSI at both the transmitter and receiver can be a strong limitation in practice. One thus usually applies lower-complexity suboptimal techniques.

3.3.2. Linear Precoding

Linear Precoding (LP) [53,54] has the receivers treat interference as noise. Users' messages are encoded independently, a beamformer is chosen for each stream, the beamformer outputs are added, and the result is transmitted through the channel. The beamforming is expressed as a linear operation $\boldsymbol{x} = \boldsymbol{P}\boldsymbol{u}$, where $\boldsymbol{u}$ is a K-dimensional vector carrying the messages of the K users, and where $\boldsymbol{P}$ is the precoding matrix. The received vector of the k-th user is

$$y_k[t] = \boldsymbol{h}_k^{\mathrm{T}}\boldsymbol{p}_k u_k[t] + \sum_{i=1, i\neq k}^{K} \boldsymbol{h}_k^{\mathrm{T}}\boldsymbol{p}_i u_i[t] + z_k[t]. \tag{3.16}$$

where $\boldsymbol{p}_i$ is the i-th column of $\boldsymbol{P}$.

Zero forcing (ZF) [55, 56] is a technique that cancels the multi-user interference, which is the second summand of (3.16), but ignores the effect of noise at the receivers. The ZF precoder is the pseudo-inverse of $\boldsymbol{H}$ multiplied by the inverse of a constant α_{ZF} chosen to satisfy the power constraint:

$$\boldsymbol{P}_{ZF} = \frac{1}{\alpha_{\mathrm{ZF}}}\boldsymbol{H}^{\mathrm{H}}(\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}})^{-1}, \;\; \alpha_{\mathrm{ZF}} = \sqrt{\mathrm{tr}\left((\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}})^{-1}\right)}. \tag{3.17}$$

The matched filter (MF) is designed to maximize the single-user SNR, and therefore in the multi-user system the precoder ignores multi-user interference. For a frequency selective channel the linear precoding matrix and the normalization coefficient are computed per subcarrier in the frequency domain:

$$\boldsymbol{P}_{MF} = \frac{1}{\alpha_{\mathrm{MF}}}\boldsymbol{H}^{\mathrm{H}}, \;\; \alpha_{\mathrm{MF}} = \sqrt{\mathrm{tr}\,(\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}})}. \tag{3.18}$$

Finally, the linear minimum mean square error (MMSE) [56] precoder minimizes the squared norm of the distortion between the receiver estimate of the signal and the transmitted signal:

$$\begin{aligned}\boldsymbol{P}_{WF} &= \frac{1}{\alpha_{\mathrm{WF}}}\boldsymbol{H}^{\mathrm{H}}\left(\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}} + \frac{1}{K\cdot SNR}\boldsymbol{I}\right)^{-1}\\ &= \frac{1}{\alpha_{\mathrm{WF}}}\boldsymbol{P}\\ \alpha_{\mathrm{WF}} &= \sqrt{\mathrm{tr}\,((\boldsymbol{P}\boldsymbol{P}^{\mathrm{H}}))}.\end{aligned} \tag{3.19}$$

This precoder, also called a Wiener Filter (WF), maximizes the received Signal to Interference Noise Ratio (SINR) under symmetry assumptions [57].

3.3.3. Information Rate Lower Bounds

Capacity is difficult to approach because of the discrete alphabet constraints. We instead consider mutual information lower bounds for single user decoding. To establish lower bounds, we model the per-user channels as equivalent parallel channels as illustrated in Fig. 3.2. The intuition follows from the literature on massive MIMO as well as interference channels [58–61] which shows that treating interference as additive noise is close to optimal in the weak interference regime.

We use the mismatched decoding framework [62, 63], see Appendix A. Consider a channel $p_{Y|X}(\cdot|\cdot)$ with input X and output Y. A lower bound on the mutual information

$$\mathrm{I}(X;Y) = \mathrm{E}\left[\log_2\left(\frac{p_{Y|X}(Y|X)}{\sum_a p_{Y|X}(Y|a)P_X(a)}\right)\right] \tag{3.20}$$

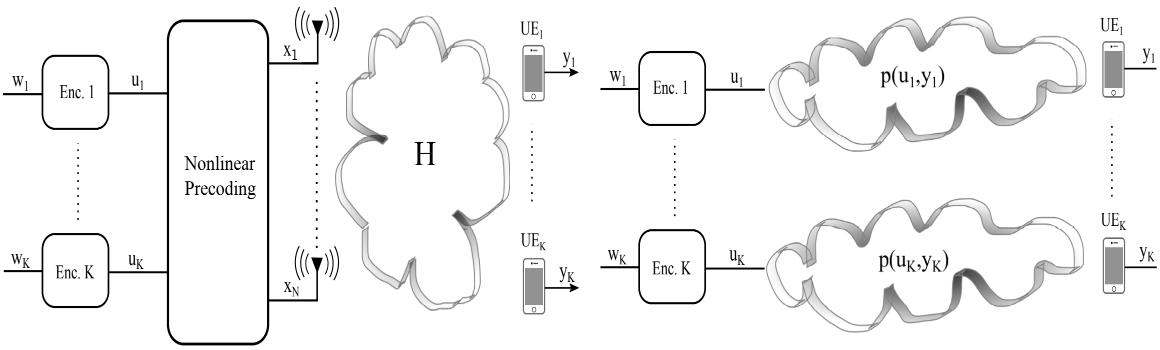


Figure 3.2.: Non-linear Precoding (left) and Equivalent Interpretation (right)

is the generalized mutual information (GMI)

$$\mathrm{E}\left[\log_2\left(\frac{q_{Y|X}(Y|X)^s}{\sum_a q_{Y|X}(Y|a)^s P_X(a)}\right)\right] \tag{3.21}$$

where $q_{Y|X}(\cdot|\cdot)$ is an *auxiliary* channel. The lower bound is tight if $q_{Y|X} = p_{Y|X}$. However, $p_{Y|X}$ is often difficult to characterize and hence we resort to tractable channels $q_{Y|X}$ to simplify the calculations and get insight on receiver design.

To benchmark the performance of different precoding strategies, consider Scenario 1 where the receivers know the α values. As mentioned above, this scenario may be unrealistic, but we will see that the performance trends for Scenarios 2 and 3 are similar. We proceed as follows:

1. Perform Monte Carlo simulations to collect long sample sequences $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(S)}$ and $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(S)}$ of length S from the true channel.

2. We distinguish different cases.

   ▷ For Scenario 1 in Sec. 3.2.3, the receiver knows the α values, and it scales each received sample with the corresponding α, i.e., we compute

   $$\tilde{\boldsymbol{y}}^{(i)} = \alpha^{(i)}\boldsymbol{y}^{(i)}, \quad i = 1, \ldots, S. \tag{3.22}$$

   ▷ For Scenarios 2 and 3, the receivers do not know α. Instead, the effect of the α values is captured in the estimation of the channel coefficient, see step 3 below, and we set $\tilde{\boldsymbol{y}}^{(i)} = \boldsymbol{y}^{(i)}$. Note that, because the precoding factor depends on $\boldsymbol{u}$, the estimated α are inherently mismatched, as only pilot symbols are available to the receiver.

▷ For the frequency selective case of problem (3.11), the precoding factor is updated once per coherence interval. As before, we have $\tilde{\boldsymbol{y}}[t]^{(i)} = \boldsymbol{y}[t]^{(i)}, t = 1, \ldots, T$, and the calculated channel in step 3 below is used for all received samples within one coherence interval.

3. Every receiver chooses a Gaussian auxiliary channel $\tilde{Y} = h \cdot U + Z$ with conditional density

$$q_{\tilde{Y}|U}(\tilde{y}|u; h, \sigma_q^2) = \frac{1}{\pi\sigma_q^2}\, \mathrm{e}^{-\frac{|\tilde{y}-h\cdot u|^2}{\sigma_q^2}}. \tag{3.23}$$

The parameters $h \in \mathbb{C}$ and $\sigma_q^2 \in \mathbb{R}^+$ are obtained by maximum-likelihood (ML) estimation from the sample sequences for a particular user k:

$$h = \frac{\sum_{i=1}^{S} \tilde{y}_k^{(i)} u_k^{(i)*}}{\sum_{i=1}^{S} \left|u_k^{(i)}\right|^2}; \qquad \sigma_q^2 = \frac{1}{S}\sum_{i=1}^{S} \left|\tilde{y}_k^{(i)} - h u_k^{(i)}\right|^2. \tag{3.24}$$

A Gaussian auxiliary channel seems reasonable, especially for large N.

4. Estimate the GMI as the empirical mean

$$R_\mathrm{a} \approx \max_{s\geq 0} \frac{1}{S}\sum_{i=1}^{S} \log_2\left(\frac{q_{\tilde{Y}|U}\left(\tilde{y}_k^{(i)}|u_k^{(i)}\right)^s}{\sum_{a\in\mathcal{U}} q_{\tilde{Y}|U}\left(\tilde{y}_k^{(i)}|a\right)^s \frac{1}{|\mathcal{U}|}}\right). \tag{3.25}$$

Fig 3.3 shows the calculated R_a for a massive MIMO scenario with $N = 128$ antennas at the base station and a considerably lower $K = 16$ number of served users. Note that the capacity with uniform power allocation and the WF precoder lower bound are close, meaning that linear precoding with Gaussian inputs achieves rates that are close to the capacity. Furthermore, with uniformly distributed 256−QAM and the lower bound (3.25), the WF operates within the shaping gap from the Gaussian lower bound. Shaping will be needed to close this gap [64].

As outlined above, GMI can address general input distributions, metrics that incorporate channel estimation, and sub-optimal decoders. An educated guess for the

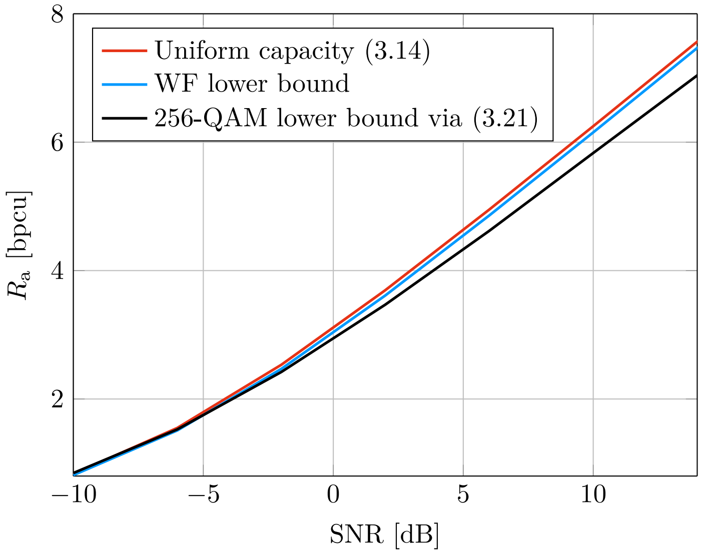


Figure 3.3.: Rate per user vs. SNR for $N = 128$ and $K = 16$.

auxiliary channel $q_{Y|X}$ also gives insight because the rates are achievable by a decoder performing maximum likelihood decoding using the chosen metric [62]. On the other hand, the method provides suboptimal rates and the choice of $q_{Y|X}$ may hide relevant problem structure. For example, our choice of metric is similar to the nearest neighbor decoding rule for Additive White Gaussian Noise (AWGN) channels. This means that our decoder treats distortion as Gaussian and thus opts for a worst case assumption on the noise.

In the following section we present alternative lower bounds that can be seen as particular instances of the GMI with different implicit or explicit choices of the auxiliary channel.

3.3.4. Lower Bounds with a Gaussian Noise Model

A memoryless channel model for a Gaussian interference channel is given by:

$$y_k[t] = c_k u_k[t] + \sum_{i=1, i\neq k}^{K} c_i u_i[t] + z_k[t] \tag{3.26}$$

where the channel coefficients c_k and c_i for users k and i, respectively, specify the joint effect of the channel and the precoder.

We collect these coefficients in the vector $\boldsymbol{c}$. The mutual information of the channel input and output given $\boldsymbol{C} = \boldsymbol{c}$ is:

$$\begin{aligned}
\mathrm{I}(U_k; Y_k|\boldsymbol{C} = \boldsymbol{c}) &= h(U_k) - h(U_k|Y_k, \boldsymbol{C} = \boldsymbol{c}) \\
&= h(U_k) - h(U_k - \hat{U}_k|Y_k, \boldsymbol{C} = \boldsymbol{c}) \\
&\geq h(U_k) - \log_2 \operatorname{var}(U_k - \hat{U}_k|Y_k, \boldsymbol{C} = \boldsymbol{c}) \\
&= \log_2(\pi e) - \log_2\left(\pi e \frac{\sigma^2 + \sum_{i=1,i\neq k}^{K} |c_i|^2}{|c_k|^2 + \sigma^2 + \sum_{i=1,i\neq k}^{K} |c_i|^2}\right) \\
&= \log_2\left(1 + \frac{|c_k|^2}{\sigma^2 + \sum_{i=1,i\neq k}^{K} |c_i|^2}\right) \qquad (3.27)
\end{aligned}$$

where we applied the maximum entropy theorem and a Gaussian distribution of the transmitted symbols with $P = 1$. The ergodic rate is defined as the average mutual information with respect to the distribution of the channel coefficients:

$$R_k = I(U_k; Y_k|\boldsymbol{C}) = \mathrm{E}\left[\log_2\left(1 + \frac{|C_k|^2}{\sigma^2 + \sum_{i=1,i\neq k}^{K} |C_i|^2}\right)\right]. \qquad (3.28)$$

Another class of lower bounds is based on treating quantization noise as additive noise that is uncorrelated with the input; this is known as a Bussgang decomposition [65, Thm. 2]. According to this decomposition, the transmitted vector $\boldsymbol{x}$ can be written as:

$$\boldsymbol{x}[t] = \boldsymbol{G}\boldsymbol{P}\boldsymbol{u}[t] + \boldsymbol{d}[t]. \qquad (3.29)$$

where $\boldsymbol{P}$ is the precoding matrix, $\boldsymbol{G}$ is a diagonal matrix that is a nonlinear function of $\boldsymbol{P}$, and $\boldsymbol{d}[t]$ is noise where $\mathrm{E}\left[\boldsymbol{u}[t]\boldsymbol{d}[t]^{\mathrm{H}}\right] = 0$. The received signal for user k is now

$$y_k[t] = c_k u_k[t] + \sum_{i\neq k} c_i u_i[t] + (\boldsymbol{h}_k^{\mathrm{T}}\boldsymbol{d}[t] + z_k[t]) \qquad (3.30)$$

where $c_k = \boldsymbol{h}_k^{\mathrm{T}}\boldsymbol{G}\boldsymbol{p}_k$ and $c_i = \boldsymbol{h}_k^{\mathrm{T}}\boldsymbol{G}\boldsymbol{p}_i$.

The authors of [65] propose a lower bound based on (3.27) by assuming that the total noise term is Gaussian with the same covariance as $\sum_{i\neq k} c_i u_i[t] + (\boldsymbol{h}_k^{\mathrm{T}} \boldsymbol{d}[t] + z_k[t])$. $\boldsymbol{G}$ and the covariance matrix $\boldsymbol{C}_d = \mathrm{E}\left[\boldsymbol{d}\boldsymbol{d}^{\mathrm{H}}\right]$ can be numerically computed or approximated using asymptotic results from random matrix theory [65]. The resulting approximation is

$$\mathrm{I}(U_k; Y_k|\boldsymbol{C} = \boldsymbol{c}) \approx \log_2 \left(1 + \frac{|c_k|^2}{\sigma^2 + \boldsymbol{h}_k^{\mathrm{T}} \boldsymbol{C}_d \boldsymbol{h}_k^{\star} + \sum_{i=1, i\neq k}^{K} |c_i|^2}\right) \tag{3.31}$$

We remark that, as argued above, (3.31) is a valid lower bound when the U_k are iid and Gaussian with variance $P = 1$.

3.4. Non-linear Precoding Algorithms

This section reviews algorithms that address the symbol-wise precoding problem for discrete signaling. We begin with a literature review.

3.4.1. Prior Work

There are numerous studies of the *uplink* with either linear detectors, e.g., Matched Filter (MF), Zero Forcing (ZF) and Wiener Filter (WF), or non-linear detectors such as Approximate Message Passing (AMP) [66–68]. For example, even for low-resolution quantization at the UEs, reliable communication with higher order modulation is possible if N is sufficiently large [68].

The *downlink* has also received attention [65, 69–76]. The authors of [69] use the Bussgang decomposition to design Quantized Linear Precoding (QLP). The paper [70] introduces a lookup-table based precoder for Quadrature Phase Shift Keying (QPSK) that minimizes the uncoded Bit Error Rate (BER) at the UE. The paper [71] introduces a hybrid RF architecture that combines a constrained and a conventional MIMO array. A greedy knapsack-like algorithm is used to minimize the mean square error (MSE) between the desired and constructed UE signal points.

The authors of [65] describe nonlinear approaches based on semidefinite relaxation, ℓ_∞ norm relaxation and sphere decoding. For a reasonable performance and complexity tradeoff, they recommend ℓ_∞ norm relaxation, which is named SQUared-Infinity norm Douglas-rachford splitting (SQUID) [65, Sec. IV]. Another approach is described in [72], where the authors extend the framework of Alternating Direction Method of Multipliers (ADMM), and they report slight improvements over SQUID. The precoding problem for coarsely quantized, frequency-selective channels and its integration with OFDM was considered in [73], where the authors use linear precoding and a frequency domain approach. An extension of SQUID to OFDM and frequency selective channels was presented in [42]. A different approach to symbol level precoding is taken in [74], where the authors propose a scheme that starts from a minimum symbol error probability and relaxes the discrete set constraint to obtain a linear program. The solution of the linear program is projected element-wise on the discrete set to obtain the precoding vector. The authors from the same research group extend this method to frequency selective channels in [75].

We proposed an iterative algorithm called Multiple Antenna Greedy Iterative Quantized (MAGIQ) in [76] inspired by coordinate minimization. A simple extension of MAGIQ is here called Quantized Coordinate Minimization (QCM). These algorithms are equally well-suited for flat and frequency selective channels and offer an advantageous trade-off between simplicity, computational efficiency and achievable rates. An algorithm similar to MAGIQ/QCM was proposed in [77] at about the same time. These algorithms use a cyclic coordinate descent algorithm to perform the OFDM symbol-wise precoding, and they use MSE as a cost function. The same group [78] proposed a hardware architecture and precoding algorithm that are similar to the ones presented in [76].

Another approach is described in [79], where the authors derive bounds for nonlinear precoding based on MSE minimization. Replica method [80] approximations are shown to be give inexact predictions for discrete signaling, and they require sophisticated approaches such as 1- or 2-replica symmetry breaking to tighten the approximation.

3.4.2. Quantized Linear Precoding

A heuristic approach to quantized precoding is to apply a scalar quantizer to a LP output. Quantized Linear Precoding (QLP) approximates the solution of (3.9) by $\boldsymbol{x} = \mathsf{Q}(\boldsymbol{P}\boldsymbol{u})$, where $\boldsymbol{P} \in \mathbb{C}^{N\times K}$ is a precoding matrix and $\mathsf{Q}(\cdot)$ is a quantization function with range $\mathcal{X}$ that operates entry-wise on $\boldsymbol{x}$. QLP is conceptually simple, and inherits the computational complexity advantages of linear precoding. However, it performs poorly for higher order modulations and intermediate ranges of K/N, as shown below by both simulation and analysis.

3.4.3. SQUID Algorithm

Define the auxiliary variable $\boldsymbol{z} = \alpha\boldsymbol{x}$ and rewrite (3.9) in the following form:

$$\underset{\boldsymbol{z}\in\mathcal{B}^N}{\text{minimize}} \quad \|\boldsymbol{u} - \boldsymbol{H}\boldsymbol{z}\|_2^2 + \frac{K\sigma^2}{P}\,\|\boldsymbol{z}\|_2^2 \tag{3.32}$$

where $\mathcal{B} = \alpha\mathcal{X}$ and $\alpha^2 = ||\boldsymbol{z}||_2^2/P$. The complex-valued problem can be transformed into a real-valued one by the following transformations:

$$\boldsymbol{z}_{\mathbb{R}} = \begin{bmatrix} \Re\{\boldsymbol{z}\} \\ \Im\{\boldsymbol{z}\} \end{bmatrix}, \ \boldsymbol{u}_{\mathbb{R}} = \begin{bmatrix} \Re\{\boldsymbol{u}\} \\ \Im\{\boldsymbol{u}\} \end{bmatrix}, \text{ and } \boldsymbol{H}_{\mathbb{R}} = \begin{bmatrix} \Re\{\boldsymbol{H}\} & -\Im\{\boldsymbol{H}\} \\ \Im\{\boldsymbol{H}\} & \Re\{\boldsymbol{H}\} \end{bmatrix}. \tag{3.33}$$

Now (3.32) can be reformulated as a real valued problem that can be relaxed by replacing the discrete set constraint with $\boldsymbol{z}_{\mathbb{R}} \in \mathbb{R}$. We thus have

$$\begin{aligned} &\underset{\boldsymbol{z}_{\mathbb{R}}}{\text{minimize}} && \|\boldsymbol{u}_{\mathbb{R}} - \boldsymbol{H}_{\mathbb{R}}\boldsymbol{z}_{\mathbb{R}}\|_2^2 + \frac{K\sigma^2}{P}\,\|\boldsymbol{z}_{\mathbb{R}}\|_2^2 \\ &\text{subject to} && z_{\mathbb{R},1}^2 = z_{\mathbb{R},b}^2 \quad \text{for } b = 2,\ldots,2N. \end{aligned} \tag{3.34}$$

This particular form of the optimization problem is non-convex because of the equality constraint.

The authors of [65] propose to relax (3.34) by using a l_∞ regularization, and then dropping the non-convex constraint. The intent is to enforce the equality constraint through a l_∞ regularizer, much as the l_1 proves to be a good proxy for sparsity in compressed sensing. We arrive at

$$\underset{\boldsymbol{z}_{\mathbb{R}}}{\text{minimize}} \quad \|\boldsymbol{u}_{\mathbb{R}} - \boldsymbol{H}_{\mathbb{R}}\boldsymbol{z}_{\mathbb{R}}\|_2^2 + \frac{2NK\sigma^2}{P} \|\boldsymbol{z}_{\mathbb{R}}\|_{\infty}^2 . \tag{3.35}$$

The resulting expression is further split as a sum of two convex functions

$$g(\boldsymbol{z}_{\mathbb{R}}) = \|\boldsymbol{u}_{\mathbb{R}} - \boldsymbol{H}_{\mathbb{R}}\boldsymbol{z}_{\mathbb{R}}\|_2^2$$

and

$$f(\boldsymbol{z}_{\mathbb{R}}) = \frac{2NK\sigma^2}{P} \|\boldsymbol{z}_{\mathbb{R}}\|_{\infty}^2 .$$

The Douglas-Rachford Splitting (DRS) algorithm can be used to find the solution to the relaxed problem. We define the proximal operator for the function $h(\cdot)$ as

$$\text{prox}_h(\boldsymbol{w}) = \underset{\boldsymbol{z}_{\mathbb{R}}}{\text{argmin}}\, h(\boldsymbol{z}_{\mathbb{R}}) + \tfrac{1}{2}\|\boldsymbol{z}_{\mathbb{R}} - \boldsymbol{w}\|_2^2. \tag{3.36}$$

The proximal operator is a small optimization problem and converts the minimization of the function $h(\cdot)$, that is possibly not smooth nor differentiable, into a smooth and differentiable problem. The proximal operator has a number of interpretations, and an intuitive one is to consider it to be a mixture of a generalized projection on a set (the constrained domain of $h(\cdot)$) and a gradient step towards the minimum of the function.

For example, for our optimization problem the proximal operator of $g(\cdot)$ is an unconstrained smooth differentiable convex program. Thus, the minimum is found by solving

$$\frac{\mathrm{d}g(\boldsymbol{z}_{\mathbb{R}})}{\mathrm{d}\boldsymbol{z}_{\mathbb{R}}} = -2\boldsymbol{H}_{\mathbb{R}}^T(\boldsymbol{u}_{\mathbb{R}} - \boldsymbol{H}_{\mathbb{R}}\boldsymbol{z}_{\mathbb{R}}) + (\boldsymbol{z}_{\mathbb{R}} - \boldsymbol{w}) = 0 \tag{3.37}$$

$$\Rightarrow \boldsymbol{z}_{\mathbb{R}} = (\boldsymbol{H}_{\mathbb{R}}{}^{\mathrm{T}}\boldsymbol{H}_{\mathbb{R}} + \frac{1}{2}\boldsymbol{I})^{-1}(\boldsymbol{H}_{\mathbb{R}}{}^{\mathrm{T}}\boldsymbol{u}_{\mathbb{R}} + \frac{\boldsymbol{w}}{2}). \tag{3.38}$$

The proximal operator for $f(\cdot)$ does not have a closed form, but it can be computed with a simple iterative algorithm given in [65].

We now step back and derive the DRS algorithm. Define the reflection operator $R_h(x) = 2\text{prox}_h(x) - x$ and note that the proximal operator and the resolvent relation

$(\boldsymbol{I}+\partial h)$ have the same unique solution [81]. Starting from the optimality condition (from subgradient calculus) we can obtain a fixed point equation to calculate the minimum of the sum of two operators:

$$\begin{aligned}
\mathbf{0} &\in \partial f(\boldsymbol{x}) + \partial g(\boldsymbol{x}) \quad \text{optimality condition}\\
2\boldsymbol{x} &\in (\boldsymbol{I}+\partial f)(\boldsymbol{x}) + (\boldsymbol{I}+\partial g)(\boldsymbol{x})\\
2\boldsymbol{x} &\in (\boldsymbol{I}+\partial f)(\boldsymbol{x}) + \boldsymbol{z}\\
\boldsymbol{x} &= \operatorname{prox}_g(\boldsymbol{z}), \text{ the prox operator has same solution as the resolvent}\\
R_g(\boldsymbol{z}) &= (\boldsymbol{I}+\partial f)(\boldsymbol{x}) \to \boldsymbol{x} = \operatorname{prox}_f(R_g(\boldsymbol{z}))\\
\boldsymbol{z} &= 2\boldsymbol{x} - R_g(\boldsymbol{z}) = 2\operatorname{prox}_f(R_g(\boldsymbol{z})) - R_g(\boldsymbol{z}).
\end{aligned} \tag{3.39}$$

Finally, SQUID can be summarized as a fixed point iteration:

$$\begin{cases}
\boldsymbol{x} & = \operatorname{prox}_g(\boldsymbol{z})\\
\boldsymbol{z} & = R_f(R_g(\boldsymbol{z}))\\
\boldsymbol{z} & = \frac{1}{2}\boldsymbol{z} + \frac{1}{2}R_f(R_g(\boldsymbol{z})).
\end{cases} \tag{3.40}$$

The DRS fixed point procedure can be written in several ways by splitting the fixed point equation and introducing dummy variables:

$$\begin{aligned}
\boldsymbol{a}_{\mathbb{R}}^{(i)} &= \operatorname{prox}_g(2\boldsymbol{z}_{\mathbb{R}}^{(i-1)} - \boldsymbol{c}_{\mathbb{R}}^{(i-1)})\\
\boldsymbol{z}_{\mathbb{R}}^{(i)} &= \operatorname{prox}_f(\boldsymbol{c}_{\mathbb{R}}^{(i-1)} - \boldsymbol{a}_{\mathbb{R}}^{(i)} - \boldsymbol{z}_{\mathbb{R}}^{(i-1)})\\
\boldsymbol{c}_{\mathbb{R}}^{(i)} &= \boldsymbol{c}_{\mathbb{R}}^{(i-1)} + \boldsymbol{a}_{\mathbb{R}}^{(i)} - \boldsymbol{z}_{\mathbb{R}}^{(i-1)}
\end{aligned} \tag{3.41}$$

where i denotes the iteration number. These equations are iterated until either $\boldsymbol{z}_{\mathbb{R}}$ converges or a given number of steps is completed. The final solution is quantized with $\mathcal{Q}(\cdot)$ to give an approximate solution of the original problem.

3.4.4. ADMM Algorithm

To cast problem (3.9) in the ADMM framework, we formulate it as a consensus problem:

$$\begin{aligned} \min_{\boldsymbol{x}_1,\boldsymbol{x}} \quad & \|\boldsymbol{u}-\boldsymbol{H}\boldsymbol{x}_1\|_2^2 + I_{\mathcal{X}}(\boldsymbol{x}), \\ \text{s.t.} \quad & \boldsymbol{x}_1 - \boldsymbol{x} = 0 \end{aligned} \tag{3.42}$$

where $I_{\mathcal{X}}(\cdot)$ is the indicator function of $\mathcal{X}^N$:

$$I_{\mathcal{X}}(\boldsymbol{x}) = \begin{cases} 0, & \text{if } \boldsymbol{x} \in \mathcal{X}^N, \\ \infty, & \text{otherwise.} \end{cases} \tag{3.43}$$

The augmented Lagrangian of (3.42) is expressed as

$$L_r\left(\boldsymbol{x}_1, \boldsymbol{x}, \boldsymbol{s}\right) = \|\boldsymbol{u}-\boldsymbol{H}\boldsymbol{x}_1\|_2^2 + I_{\mathcal{X}}(\boldsymbol{x}) + \boldsymbol{s}^H(\boldsymbol{x}_1 - \boldsymbol{x}) + \gamma\left\|\boldsymbol{x}_1 - \boldsymbol{x}\right\|_2^2 \tag{3.44}$$

where $\boldsymbol{s}$ is the dual vector, and $\gamma > 0$ is a penalty parameter (or the augmented Lagrangian parameter). The ADMM for this problem is:

$$\begin{aligned} \boldsymbol{x}_1^{(i+1)} &= \left(\boldsymbol{H}^H\boldsymbol{H} + \gamma\boldsymbol{I}\right)^{-1}\left(\boldsymbol{H}^H\boldsymbol{u} + \gamma\boldsymbol{x}^{(i)} - \boldsymbol{s}^{(i)}\right) \\ \boldsymbol{x}^{(i+1)} &= \Pi_{\mathcal{X}}\left(\boldsymbol{x}_1^{(i+1)} + \frac{1}{2\gamma}\boldsymbol{s}^{(i)}\right) \\ \boldsymbol{s}^{(i+1)} &= \boldsymbol{s}^{(i)} + \gamma\left(\boldsymbol{x}_1^{(i+1)} - \boldsymbol{x}^{(i+1)}\right) \end{aligned} \tag{3.45}$$

where $\Pi_{\mathcal{X}}$ is the entry-wise projection onto $\mathcal{X}$. In (3.45), the $\boldsymbol{x}_1$-update solves a minimization problem, the $\boldsymbol{x}$-update involves projection onto the finite-alphabet set $\mathcal{X}^N$, and the $\boldsymbol{s}$-update can be interpreted as a consensus adjustment step with step size γ.

It is known [72] that the ADMM algorithm may not converge when applied to non-convex problems. Unfortunately this happens with the formulation in (3.45). It turns out that the projection onto the finite set $\mathcal{X}$ is the main culprit, and to fix this the authors of [72] dampen the updates so that there is no oscillatory behavior or sudden transitions from one iteration to the next. Furthermore, the dual variable update also results in divergent behavior, so it is eliminated. The result is the expression:

$$\begin{aligned}\boldsymbol{x}_1^{(i+1)} &= \left(\boldsymbol{H}^H\boldsymbol{H}+\gamma\boldsymbol{I}\right)^{-1}\left(\boldsymbol{H}^H\boldsymbol{u}+\gamma\boldsymbol{x}^{(i)}\right)\\ \boldsymbol{x}^{(i+1)} &= \Pi_{\mathcal{X}}\left(\boldsymbol{x}_1^{(i+1)}\right)\\ \boldsymbol{x}^{(i+1)} &= \beta\boldsymbol{x}^{(i)}+(1-\beta)\boldsymbol{x}^{(i+1)}.\end{aligned} \tag{3.46}$$

This algorithm still does not to converge and it is observed that the update step for $\boldsymbol{x}_1^{(i+1)}$ is a biased MMSE estimator, and steps are taken to debias the estimation at least asymptotically as $N \to \infty$. The matrix $\left(\boldsymbol{H}^H\boldsymbol{H}+\gamma\boldsymbol{I}\right)^{-1}\boldsymbol{H}^H$ is thus augmented as $\boldsymbol{D}\left(\boldsymbol{H}^H\boldsymbol{H}+\gamma\boldsymbol{I}\right)^{-1}\boldsymbol{H}^H$ where the matrix $\boldsymbol{D}$ is chosen so that the diagonal values of the product are 1, and the off diagonals vanish.

Since the values of γ tend to be large one can further approximate:

$$(\boldsymbol{H}^H\boldsymbol{H}+\gamma\boldsymbol{I})^{-1} \approx \frac{1}{\gamma}\boldsymbol{I}. \tag{3.47}$$

We use this simplified algorithm as a benchmark and we refer to it as ADMM from here on:

$$\begin{aligned}\boldsymbol{x}^{(i+1)} &= \Pi_{\mathcal{X}}\left(\boldsymbol{x}^{(i)}+\boldsymbol{D}\left(\boldsymbol{u}-\boldsymbol{H}\boldsymbol{x}^{(i)}\right)\right)\\ \boldsymbol{x}^{(i+1)} &= \beta\boldsymbol{x}^{(i)}+(1-\beta)\boldsymbol{x}^{(i+1)}.\end{aligned} \tag{3.48}$$

3.4.5. MSM Algorithm

The Maximum Safety Margin (MSM) algorithm [74] is based on the idea of exploiting interference. Constructive interference starts from the observation that only the interference that moves a received symbol close to the border of, or beyond, the detection region is disadvantageous and should be controlled. Precoding has to therefore jointly consider the transmitted symbol, the channel, the quantizer and AWGN noise power.

The symbols to be reconstructed at the receivers belong to the M-PSK constellation defined by the set $\mathbb{U}$:

$$\mathbb{U} := \left\{e^{\mathrm{j}(2i-1)\theta} : i = 1,\cdots,M\right\},\ \text{where } \theta = \frac{\pi}{M}. \tag{3.49}$$

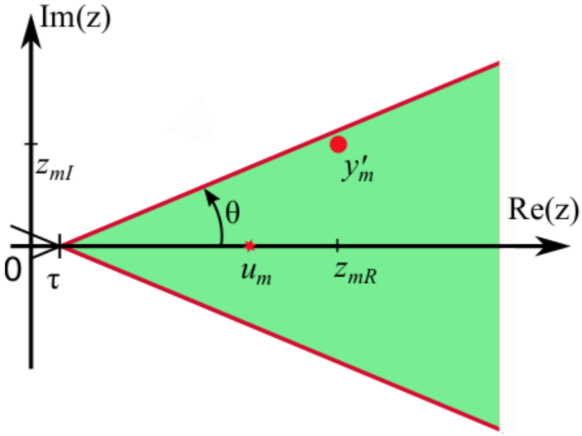


Figure 3.4.: PSK detection cell in a rotated coordinate system.

The Voronoi detection cell determined by $\mathbb{U}$ in the complex plane is shown as the shaded green area in Fig. 3.4, rotated so that the elements of $\mathbb{U}$ are brought on the real axis. This formulation allows to state the optimization problem independent of angle.

The complex noiseless received signal is given by $\boldsymbol{y\prime} = \boldsymbol{Hx}$, and the vector of coordinates on the real and imaginary axes of the rotated coordinate system are given by $\boldsymbol{z}_R = \Re\{\boldsymbol{Hx} \circ \boldsymbol{u}^*\}$ and $\boldsymbol{z}_I = \Im\{\boldsymbol{Hx} \circ \boldsymbol{u}^*\}$, respectively, where '∘' stands for element-wise multiplication. The shaded region is uniquely described by the pair of equations:

$$|\Im\{\boldsymbol{Hx} \circ \boldsymbol{u}^*\}| \leq (\Re\{\boldsymbol{Hx} \circ \boldsymbol{u}^*\} - \tau\mathbb{1}_M) \tan\theta \tag{3.50}$$

$$\Re\{\boldsymbol{Hx} \circ \boldsymbol{u}^*\} \geq \tau.\mathbb{1}_M \tag{3.51}$$

The authors in [74] adopt a real-valued decomposition which further casts the optimization as a problem over the reals:

$$\Re\{\tilde{\boldsymbol{H}}\boldsymbol{x}\} = \underbrace{\begin{bmatrix} \Re\{\tilde{\boldsymbol{H}}\} & -\Im\{\tilde{\boldsymbol{H}}\} \end{bmatrix}}_{=\boldsymbol{A}} \underbrace{\begin{bmatrix} \Re\{\boldsymbol{x}\} \\ \Im\{\boldsymbol{x}\} \end{bmatrix}}_{=\boldsymbol{x}'} = \boldsymbol{Ax}' \tag{3.52}$$

$$\Im\{\tilde{\boldsymbol{H}}\boldsymbol{x}\} = \underbrace{\begin{bmatrix}\Im\{\tilde{\boldsymbol{H}}\} & \Re\{\tilde{\boldsymbol{H}}\}\end{bmatrix}}_{=\boldsymbol{B}} \begin{bmatrix}\Re\{\boldsymbol{x}\} \\ \Im\{\boldsymbol{x}\}\end{bmatrix} = \boldsymbol{B}\boldsymbol{x}' \tag{3.53}$$

where $\tilde{\boldsymbol{H}} = \mathrm{diag}(\boldsymbol{u}^*)\boldsymbol{H}$, and $\mathrm{diag}(\cdot)$ forms a diagonal matrix from a column vector.

Finally, the non-convex constraint on $\boldsymbol{x}$ is relaxed to include the polytope:

$$\boldsymbol{x}' \leq \frac{1}{\sqrt{2}}\mathbb{1}_{2N} \text{ and } -\boldsymbol{x}' \leq \frac{1}{\sqrt{2}}\mathbb{1}_{2N}. \tag{3.54}$$

With this, the MSM optimization can be formulated as a linear program:

$$\begin{aligned} &\max_{\boldsymbol{x}'} \ \tau \\ &\text{s.t.} \quad \begin{bmatrix}\boldsymbol{B} - \tan\theta \boldsymbol{A} & \frac{1}{\cos\theta}\mathbb{1}_M \\ -\boldsymbol{B} - \tan\theta\boldsymbol{A} & \frac{1}{\cos\theta}\mathbb{1}_M\end{bmatrix}\begin{bmatrix}\boldsymbol{x}' \\ \delta\end{bmatrix} \leq \mathbb{0}_{2M}. \\ &\text{and} \quad \boldsymbol{x}' \leq \frac{1}{\sqrt{2}}\mathbb{1}_{2N}, -\boldsymbol{x}' \leq \frac{1}{\sqrt{2}}\mathbb{1}_{2N}. \end{aligned} \tag{3.55}$$

The final solution is then quantized by a nearest-neighbour element-wise operator to the original constraint set $\mathcal{X}$.

All the algorithms presented above deal with the nonconvex constraint by applying relaxations and then casting the relaxed problem into a family of optimization algorithms aimed at reaching a good trade-off between the quality of the solution and computational efficiency, essentially proposing different global heuristics.

3.5. Quantized Coordinate Minimization Precoding

The problem stated in (3.9) is a non-convex combinatorial problem that can be transformed into a problem known to belong to a class of hard problems to solve exactly. For example, the case $\mathcal{X} = \{-1, +1\}$ can be reformulated into a quadratic binary minimization problem, which is known to belong to the class of NP-hard problems [82,83]. This, however, does not mean that the problem cannot be efficiently approximated by a simple deterministic algorithm.

As we have already seen, ADMM, SQUID and MSM are based on different relaxations of the constraint set and then recovering the feasible solution by projections. The approach is facilitated by nice properties of the cost function (e.g. convex, continuous, smooth) that guarantee the approximate solution is near the global optimum.

We now take a different approach that is based on *coordinate minimization.* This method is related to coordinate descent algorithms that have been successful in, e.g., compressed sensing [84] and date back to at least [85]. Observe that while the joint minimization over all coordinates x_i of $\boldsymbol{x}$ is expensive (in the worst case sense), the minimization of each coordinate x_i requires evaluating the function of a scalar only $|\mathcal{X}|$ times (which for our application is a small number e.g. 5 or 10). This and the large number of degrees of freedom available in the channel suggest that minimizing the cost function one coordinate a time, while holding the other coordinates fixed, will result not only in a computationally efficient algorithm, but also in a good approximation of the optimal solution. At each step we choose the best coordinate to update, i.e., a greedy approach. We later explore cyclic and random schedules for choosing the coordinates, and we compare these with the greedy approach.

We emphasize that our method is closely related to coordinate descent but we do not exploit gradient information, choosing instead to optimally solve the subproblems. The iterative nature of the algorithm and the discreteness of the constraint set seem to make a theoretic analysis of convergence a cumbersome task. Previous works on alternating minimization rely on properties of the function to be minimized and the constraint set that our problem does not share, more details can be found in Chapter 14 of [86]. What we can ensure is that the algorithm delivers a non-decreasing sequence of values, which in addition to the boundedness of the discrete support are also feasible and enjoy a coordinate-wise optimality.

In [76] we denoted the greedy version of the quantized coordinate minimization as MAGIQ and throughout this chapter we keep this name. We further denote as QCM the instances where we use the cyclic and random coordinate update schedules.

3.5.1. Flat Fading Channels

Algorithm 1 outlines MAGIQ for frequency-flat channels. The inner loop evaluates the cost function

$$F(\boldsymbol{x}, \alpha) = \|\boldsymbol{u} - \alpha \boldsymbol{H}\boldsymbol{x}\|_2^2 + \alpha^2 K \sigma^2 \tag{3.56}$$

and selects (without replacement) an antenna index n and a precoded symbol x_n from $\mathcal{X}$ such that (3.56) is minimized. That is, each iteration selects the best coordinate among the remaining coordinates to be explored, in a greedy fashion, best-first. The n-th coordinate of $\boldsymbol{x}$ at the i-th iteration is individually optimized as:

$$\begin{aligned} x_n^{(i)} &= \underset{z,\text{ with } \boldsymbol{x}_{\{1,\ldots,N\}\setminus n}\ ct.}{\operatorname{argmin}} \|\boldsymbol{u} - \alpha \boldsymbol{H}\boldsymbol{x}\|_2^2 + \alpha^2 K \sigma^2 \\ &= \left\| \boldsymbol{u} - \boldsymbol{h}_n z + \sum_{j=1}^{n-1} \boldsymbol{h}_j x_j^{(i)} + \sum_{j=n+1}^{N} \boldsymbol{h}_j x_j^{(i-1)} \right\|_2^2 + \alpha^2 K \sigma^2 \\ &= F(\boldsymbol{x}_{\{1,\ldots,n-1\}}^{(i)}, x_n, \boldsymbol{x}_{\{n+1,\ldots,N\}}^{(i-1)}). \end{aligned} \tag{3.57}$$

After computing a vector $\boldsymbol{x}$ with the procedure outlined above, the precoding factor α is chosen by setting the gradient of (3.9) with respect to α to zero and solving the resulting equation in α. α is thus a function of the channel, the target vector $\boldsymbol{u}$, as well as the precoding vector $\boldsymbol{x}$:

$$\begin{aligned} \frac{\partial F(\boldsymbol{x}, \alpha)}{\partial \alpha} &= \operatorname{Re}\{\boldsymbol{u}^{\mathrm{H}} \boldsymbol{H}\boldsymbol{x}\} + \alpha \|\boldsymbol{H}\boldsymbol{x}\|_2^2 + \alpha K \sigma^2 = 0 \\ \alpha &= \frac{\operatorname{Re}\{\boldsymbol{u}^{\mathrm{H}} \boldsymbol{H}\boldsymbol{x}\}}{\|\boldsymbol{H}\boldsymbol{x}\|_2^2 + K\sigma^2}. \end{aligned} \tag{3.58}$$

The simple gradient criteria for α follows from the observation that (3.9) is quadratic in α for a fixed $\boldsymbol{x}$. The algorithm then performs alternating updates of $\boldsymbol{x}$ and α until it reaches a predefined stopping criterion or a maximum number of iterations. Although the alternating optimization of biconvex problems is known not to converge to the global optimum in general, simulations presented in Sec. 3.10 show that the algorithm converges to a good local minimum with a small number of iterations, and that the quality of the local minimum is at least as good as the minima obtained by SQUID, ADMM or MSM [65, 72] at low SNR, and considerably better at high rates

Algorithm 1: MAGIQ for frequency-flat channels

```
Inputs: $\boldsymbol{u}$, $\boldsymbol{H}$, $\mathcal{S} = \{1, \ldots, N\}$, $err_{\min}$
Initialize: $\boldsymbol{x} = \boldsymbol{x}_{init}$, $\alpha = \alpha_{init}$
$err^{(0)} = \|\boldsymbol{u}\|^2$
for $i = 1 : I$ do
    $n = 1$
    $err^{(i)} = err^{(i-1)}$
    while $(err_n^{(i)} > err_{min}) \vee (err_n^{(i)} < err_{n-1}^{(i)}) \vee (n \leq N)$ do
        $(x_{n^\star}^\star, n^\star) = \operatorname{argmin}_{x_n \in \mathcal{X}, n \in \mathcal{S}} F(\boldsymbol{x}, \alpha)$
        $(x_1^{(i)}, \ldots, x_n^{(i)}, \ldots, x_N^{(i)})^{\mathrm{T}} = (x_1^{(i)}, \ldots, x_{n^\star}^\star, \ldots, x_N^{(i)})^{\mathrm{T}}$
        $\mathcal{S} \leftarrow \mathcal{S} \setminus \{n^\star\}$
        $err_n^{(i)} = \left\|\boldsymbol{u} - \alpha^{(i)} \boldsymbol{H} \boldsymbol{x}^{(i)}\right\|_2^2 + (\alpha^{(i)})^2 K \sigma^2$
        $n \leftarrow n + 1$
    end
    $\alpha^{(i)} = \frac{\mathrm{Re}\{\boldsymbol{u}^{\mathrm{H}} \boldsymbol{H} \boldsymbol{x}^{(i)}\}}{\left\|\boldsymbol{H} \boldsymbol{x}^{(i)}\right\|_2^2 + K\sigma^2}$
end
Output $\boldsymbol{x}, \alpha$
```

and at high SNR.

3.5.2. Quantized Precoding for OFDM

Fig. 3.1 shows how OFDM can be combined with quantized precoding. The frequency domain vector $\hat{\boldsymbol{u}}[\cdot]$ corresponding to the K users is converted to the time domain vector $\boldsymbol{u}[\cdot]$ by a length T Inverse Discrete Fourier Transform (IDFT):

$$\begin{aligned} u_k[t] &= \frac{1}{T} \sum_{m=1}^{T} \hat{u}_k[v] e^{\mathrm{j}\, 2\pi(m-1)(t-1)/T} \\ \hat{\boldsymbol{u}}[t] &= [\hat{u}_1[t], \ldots, \hat{u}_K[t]]^{\mathrm{T}} \\ \boldsymbol{u}[t] &= [u_1[t], \ldots, u_K[t]]^{\mathrm{T}} \end{aligned} \tag{3.59}$$

for $k = 1, \ldots, K$, $t = 1, \ldots, T$. For the simulations we generated the frequency domain symbols $\hat{u}_k[t]$ uniformly from a QPSK, 16-QAM, or 64-QAM regular constellation. Each UE performs single-user OFDM demodulation followed by a hard or soft decision. MAGIQ and QCM are flexible with respect to shaping constraints, constellation size, number of users, number of sub-carriers, and channel models.

For frequency selective channels, the vector $\boldsymbol{x}[t]$ of transmit symbols at time t should be chosen as a function of the transmit symbols at other time instances due to the channel memory, i.e., a choice for $\boldsymbol{x}[t]$ influences the channel output at times $t+1, t+2, \ldots, t+L-1$. However, a joint optimization over strings of length T is highly complex because of the exponential increase in the size of the constraint space $\mathcal{X}^{NT}$.

We approach the problem by splitting the joint optimization into a set of sub-problems with reduced complexity, see Algorithm 2. We perform a two-fold coordinate-wise splitting of the problem stated in (3.11). First, we solve the precoding problem for one time coordinate t at a time, starting at time 1 and ending at time T. Under this formulation, we replace the cost function (3.56) with:

$$\begin{aligned} & G(\boldsymbol{x}[1], \ldots, \boldsymbol{x}[t-1], \boldsymbol{x}[t], \boldsymbol{x}[t+1], \ldots, \boldsymbol{x}[T]) \\ & = \sum_{t=1}^{T} \left\| \boldsymbol{u}[t] - \alpha \sum_{l=0}^{L-1} \boldsymbol{H}[l]\boldsymbol{x}[t-l] \right\|_2^2 + \alpha^2 T K \sigma^2 \\ & = \sum_{t=1}^{T} \left\| \tilde{\boldsymbol{u}}[t] - \alpha \boldsymbol{H}[0]\boldsymbol{x}[t] \right\|_2^2 + \alpha^2 T K \sigma^2 \end{aligned} \tag{3.60}$$

where

$$\tilde{\boldsymbol{u}}[t] = \boldsymbol{u}[t] - \alpha \sum_{l=1}^{L-1} \boldsymbol{H}[l]\boldsymbol{x}[t-l]. \tag{3.61}$$

The last line in (3.60) has the same form as the cost function in (3.9). We again split this problem in a coordinate-wise fashion and solve it as we did in Algorithm 1 in the inner loop that processes antenna coordinates. Algorithm 2 then iterates over the frame samples until prescribed convergence criteria are met. Simulations show that the performance over frequency selective channels is close to that of the best known precoders over frequency flat channels with as little as 4-6 iterations.

Algorithm 2: MAGIQ precoding for frequency selective channels

```
Input: $\boldsymbol{H}[l]$, $\boldsymbol{u}[t]$, $t = 1, \ldots, T$
Initialization: $\boldsymbol{x}^{(0)}[t] = \boldsymbol{x}[t]_{init}$, $t = 0, \ldots, T-1$, $\alpha$=$\alpha_{init}$, $\mathcal{S} = \{1, \ldots, N\}$
for $i = 1 : I$ do
    for $t = 0 : T-1$ do
        while $\mathcal{S} \neq \varnothing$ do
            $(x^\star_{n^\star}, n^\star) = \mathrm{argmin}_{z_n \in \mathcal{X}, n \in \mathcal{S}}\, G(\boldsymbol{x}^{(i)}[0], \ldots, \boldsymbol{x}^{(i)}[t-1], \boldsymbol{z},$
             $\boldsymbol{x}^{(i-1)}[t+1], \ldots, \boldsymbol{x}^{(i-1)}[T-1])$
            $(x_1^{(i)}, \ldots, x_n^{(i)}, \ldots, x_N^{(i)})^{\mathrm{T}}[t] = (x_1^{(i)}, \ldots, x^\star_{n^\star}, \ldots, x_N^{(i)})^{\mathrm{T}}$
            $\mathcal{S} \leftarrow \mathcal{S} \setminus \{n^\star\}$
        end
        $\mathcal{S} = \{1, \ldots, N\}$
    end
    $\alpha^{(i)} = \frac{\sum_{t=0}^{T-1} \mathrm{Re}(\boldsymbol{u}[t]^{\mathrm{H}} \sum_{l=0}^{L-1} \boldsymbol{H}[l]\boldsymbol{x}^{(i)}[t-l])}{\sum_{t=0}^{T} \left\| \sum_{l=0}^{L-1} \boldsymbol{H}[l]\boldsymbol{x}^{(i)}[t-l] \right\|_2^2 + TK\sigma^2}$
end
Output: $\boldsymbol{x}[t]$, $t = 1, \ldots, T, \alpha$
```

3.5.3. Coordinate Update Policy

MAGIQ is a greedy coordinate minimization. This is in contrast to first order methods such as coordinate descent with a Gauss-Southwell rule (largest gradient magnitude) [87] which progresses in the steepest descent direction, but with no knowledge on how good the direction is. MAGIQ, however, progresses only through feasible points and the local choice is always optimal.

Our complexity analysis will show that the greedy rule demands high storage capacity and sophisticated memory addressing features, or alternatively increased computational complexity. From a real-time implementation point of view we would like to find simple rules that are competitive. For example, consider optimizing coordinates in a fixed order (or random but fixed at run-time). The set $\mathcal{T}$ is either an ordered set (a list) that we write as a vector $\mathcal{T} = [1, 2, \ldots, N]$ or a random permutation of $\mathcal{T}$. The resulting algorithms for both flat-fading and frequency selective channels are presented in Algorithms 3 and 4.

Fig. 3.5a shows achievable rates for 64-QAM with both $b = 2$ and $b = 3$ bits of phase for MAGIQ and for Algorithm 3, $N = 128$, $K = 16$. MAGIQ is run with 3

Algorithm 3: QCM precoding for frequency-flat channels

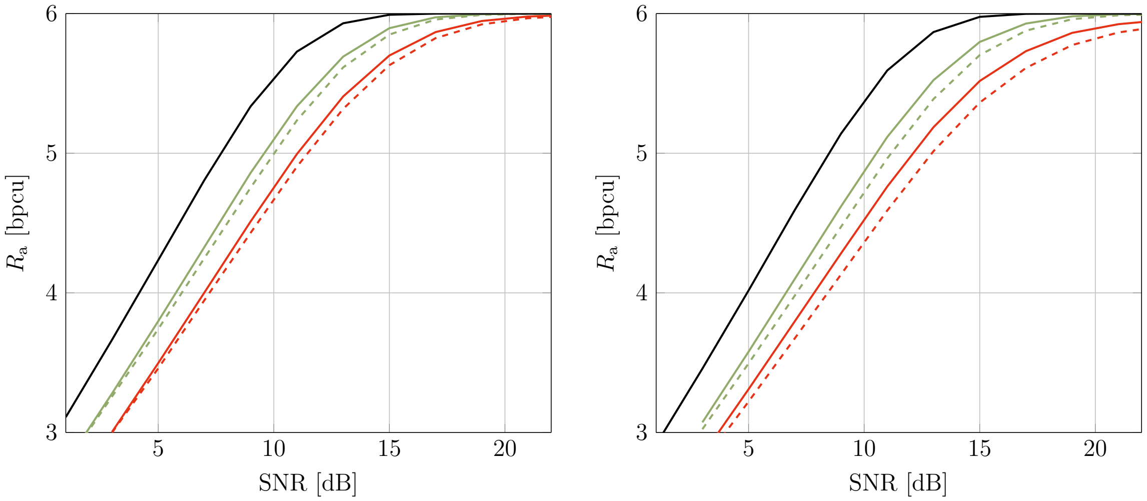
```
Inputs: $\boldsymbol{u}$, $\boldsymbol{H}$, $\mathcal{T}$, $err_{\min}$
Initialize: $\boldsymbol{x} = \boldsymbol{x}_{init}$, $\alpha = \alpha_{init}$
for $i = 1 : I$ do
    for $n \in \mathcal{T}$ do
        $x_n^\star = \operatorname{argmin}_{x_n \in \mathcal{X}} F(\boldsymbol{x}, \alpha)$
        $(x_1^{(i)}, \dots, x_n^{(i)}, \dots, x_N^{(i)})^{\mathrm{T}} = (x_1^{(i)}, \dots, x_n^\star, \dots, x_N^{(i)})^{\mathrm{T}}$
        $\mathcal{T} \leftarrow \mathcal{T} \setminus \{n\}$
    end
    $\alpha^{(i)} = \frac{\mathrm{Re}\{\boldsymbol{u}^{\mathrm{H}} \boldsymbol{H} \boldsymbol{x}^{(i)}\}}{\left\| \boldsymbol{H} \boldsymbol{x}^{(i)} \right\|_2^2 + K\sigma^2}$
end
Output $\boldsymbol{x}, \alpha$
```

Algorithm 4: QCM precoding for frequency-selective channels

```
Input: $\boldsymbol{H}[l]$, $\boldsymbol{u}[t]$, $t = 1, \dots, T$
Initialization: $\boldsymbol{x}^{(0)}[t] = \boldsymbol{x}[t]_{init}$, $t = 0, \dots, T-1$, $\alpha$=$\alpha_{init}$, $\mathcal{T}$
for $i = 1 : I$ do
    for $t = 0 : T - 1$ do
        for $n \in \mathcal{T}$ do
            $x_n^\star = \operatorname{argmin}_{z_n} G(\boldsymbol{x}^{(i)}[0], \dots, \boldsymbol{x}^{(i)}[t-1], \boldsymbol{z},$
             $\boldsymbol{x}^{(i-1)}[t+1], \dots, \boldsymbol{x}^{(i-1)}[T-1])$
            $(x_1^{(i)}, \dots, x_n^{(i)}, \dots, x_N^{(i)})^{\mathrm{T}}[t] = (x_1^{(i)}, \dots, x_n^\star, \dots, x_N^{(i)})^{\mathrm{T}}$
            $\mathcal{T} \leftarrow \mathcal{T} \setminus \{n\}$
        end
    end
    $\alpha^{(i)} = \frac{\sum_{t=0}^{T-1} \mathrm{Re}(\boldsymbol{u}[t]^{\mathrm{H}} \sum_{l=0}^{L-1} \boldsymbol{H}[l] \boldsymbol{x}^{(i)}[t-l])}{\sum_{t=0}^{T-1} \left\| \sum_{l=0}^{L-1} \boldsymbol{H}[l] \boldsymbol{x}^{(i)}[t-l] \right\|_2^2 + TK\sigma^2}$
end
Output: $\boldsymbol{x}[t]$, $t = 1, \dots, T, \alpha$
```

iterations, while Algorithm 3 is run for 5 iterations because of slower convergence. For a uncorrelated Rayleigh fading scenario the penalty for not using a greedy rule is only about 0.5dB at high SNR and almost zero at very low SNR.

In contrast, Fig. 3.5b shows the achievable rates with the transmitter correlation matrix eigenvalue profile given in Fig. 3.6. This corresponds to a linear uniform array



(a) MAGIQ vs QCM for uncorrelated Rayleigh flat fading

(b) MAGIQ vs QCM for correlated Rayleigh flat fading

Figure 3.5.: Impact of greedy search with ZF infinite resolution (——), MAGIQ $b = 2$ (——), MAGIQ $b = 3$ (——), QCM $b = 2$ (- - -), QCM $b = 3$ (- - -)

with the Jakes correlation model [88] and $\lambda/4$ antenna spacing. The receivers are assumed to be uncorrelated. Note that the gap between MAGIQ and QCM increases to $0.8 - 1.2$dB at high SNR, while it becomes non-negligible at lower SNR's. These results emphasize that the computationally complex greedy rule is more robust to less than ideal channel conditions, although by a small margin. It remains as a topic for further studies to establish the feasibility of these heuristics for more sparse channels and also with smaller numbers of transmit antennas (and implicitly degrees of freedom). We include further results for the frequency selective case in Sec. 3.10.

3.6. Complexity Analysis

This section studies the complexity of MAGIQ and QCM and compares it with SQUID [65, Sec. IV], ADMM [72] and MSM [74]. We begin with an example to show how to implement MAGIQ with a small number of complex multiplications. We emphasize that this kind of implementation may not be preferred as it complicates memory storage and addressing and may result in larger latency for real-time applications.

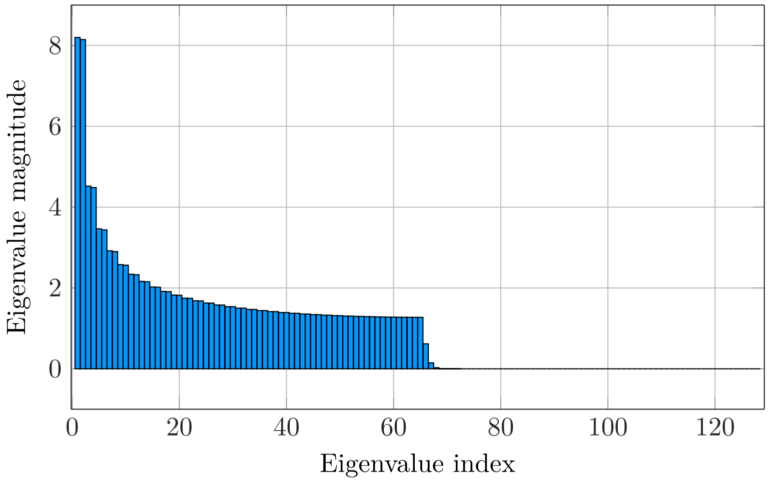


Figure 3.6.: Transmit correlation eigenvalue profile

Nonetheless, the example shows that MAGIQ offers significant flexibility.

Consider a system with 3 antennas as 2 users:

$$\boldsymbol{H} = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \end{pmatrix}.$$

For the norm calculation $||\boldsymbol{u} - \boldsymbol{H}\boldsymbol{x}||^2$ we compute and store $\boldsymbol{H}\boldsymbol{x}$ for each antenna and for all symbols in the alphabet $\mathcal{X}$ (here we can also use the symmetry of $\mathcal{X}$ and store only a quadrant and apply sign changes where appropriate). For $x_1 \in \{0, 1, -1\}$, for the first antenna we compute and store

$$\boldsymbol{H} = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \end{pmatrix} \begin{pmatrix} x_1 \\ 0 \\ 0 \end{pmatrix}.$$

For the second antenna we compute and store

$$\boldsymbol{H} = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \end{pmatrix} \begin{pmatrix} 0 \\ x_2 \\ 0 \end{pmatrix}.$$

These complex multiplications require only sign changes since the symbols in our alphabet are $\{+1, -1\}$(for 1 bit of phase) or $\{1+j, 1-j, -1+j, -1-j\}$ (for 2 bits of phase). We now compute

$$\left\| \begin{pmatrix} u_1 \\ u_1 \end{pmatrix} - \begin{pmatrix} h_{11}x_1 + h_{12}x_2 + h_{13}x_3 \\ h_{21}x_1 + h_{22}x_2 + h_{23}x_3 \end{pmatrix} \right\|^2 .$$

Since we have to do the same calculations for each coordinate of $\boldsymbol{u}$, we focus on u_1. For the first iteration of the algorithm, we compute $|u_1 - h_{1i}x_i|^2$ for $i = 1, 2, 3$:

$$\begin{aligned} |u_1 - h_{1i}x_i|^2 =& u_1 u_1^* - 2\,\mathrm{Re}\,\{h_{1i}x_i u_1^*\} + h_{1i}x_i x_i^* h_{1i}^* \\ =& \ |u_1|^2 - 2\,\mathrm{Re}\,\{h_{1i}x_i u_1^*\} + |h_{1i}|^2. \end{aligned} \tag{3.62}$$

The first and last terms in (3.62) are the absolute values squared of u_1 and h_{1i}, which can be precomputed and stored. The term $|u_1|^2$ does not depend on x_i and and can be dropped. We are thus left with NK complex multiplications. Products of the type $h_{1i}u_1^* x_i$ can be written as $h_{1i}u_1^*$ (see the discussion above about the structure of $\mathcal{X}$). We compute and store these values, and that results in another NK complex multiplications. In total so far we need $2NK$ complex multiplications and the memory storage for the results.

After the first iteration, suppose that $x_1 = +1$ on antenna 1 gives the largest reduction in MSE and is therefore fixed. We are now left with antenna 2 and antenna 3 to update. Because we have fixed antenna 1 to transmit $+1$, the updated value for u_1 is $u_1' = u_1 - h_{11}$. For antenna 2 the function that is evaluated is

$$|u_1' - h_{12}x_2|^2 = u_1' u_1'^* - 2\Re\{h_{12}x_2 u_1'^*\} + h_{12}x_2 x_2^* h_{12}^*$$

which can be expanded as:

$$\begin{aligned} &|h_{12}|^2 \ \text{ already computed} \\ &h_{12}x_2 u_1'^* = h_{12}x_2 u_1^* - h_{12}x_2 h_{11}^* = x_2(h_{12}u_1^* - h_{12}h_{11}^*). \end{aligned}$$

The term that we did not consider already for pre-computing is $h_{12}h_{11}^*$(respectively

$h_{13}h_{11}^*$ for the third antenna).

Suppose that the second antenna is selected as the best candidate with $x_2 = -1$. The updated value for u_1' is $u_1'' = u_1' + h_{12} = u_1 - h_{11} + h_{12}$. Finally we have:

$$
\begin{aligned}
&|u_1'' - h_{13}x_3|^2 = u_1''u_1''^* - 2\Re\{h_{13}x_3u_1''^*\} + h_{13}x_3x_3^*h_{13}^* \\
&|h_{13}|^2 \text{ already computed} \\
&h_{13}x_3u_1''^* = h_{13}x_3u_1^* - h_{13}x_3h_{11}^* + h_{13}x_3h_{12}^* = x_3(h_{13}u_1^* - h_{13}h_{11}^* + h_{13}h_{12}^*).
\end{aligned}
$$

The terms of the type $h_{ij}h_{ik}$ and the conjugate versions can be stored. For symmetry reasons one would need to store only ordered permutations and manage the conjugates and multiplications with x_i as sign changes. For our example we need to store $h_{11}h_{12}, h_{11}h_{13}, h_{13}h_{12}$, which means that for the general case there would be N terms to compute and store per equation. Since there are K equations, we would need an additional NK complex multiplications. Including the final NK complex multiplications from the initialization with the quantized MF solution, this results in $3NK + NK$ complex multiplications and the corresponding storage space for these terms. In practice (for example on platforms that use GPUs) it may be more efficient and convenient to perform online multiplications of terms such as $h_{13}x_3u_1''^*$, since that will save memory access and many additions and subtractions.

A worst case running time for Algorithm 1 would have the *while* loop evaluated a maximum number of N of times. Evaluating the minimum requires $\log_2(N|\mathcal{X}|)$ comparisons and therefore a worst case of $\sum_{i=1}^{N-1}\log_2((N-i)|\mathcal{X}|)$ comparisons. In addition, except for the terms already considered for multiplications, all the other operations left in computing norms and comparisons are additions and subtractions, and we will not include them in our complexity analysis.

The worst case complexity seems large, but the average complexity of MAGIQ can be substantially reduced. For example, one can initialize the algorithm with a good starting point $\boldsymbol{x}_{init}$. The results presented below are based on initializing MAGIQ with the QLP solution of the MF, which adds only NK multiplications. One can further update a coordinate only if the reduction of the MSE is significant. The significance level can be translated into a threshold for the update.

For example, Fig. 3.7 shows the empirical Cumulative Distribution Function (CDF)

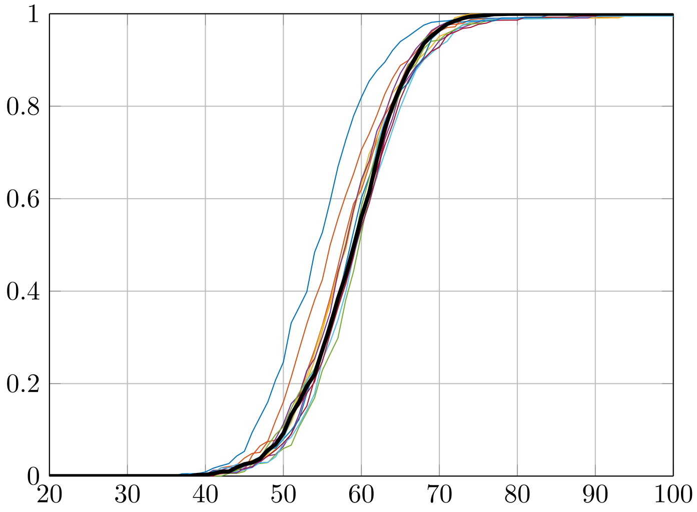


Figure 3.7.: Empirical CDF of the number of iterations with $N = 128$ and $I = 6$. The average CDF is the thick black curve.

of the total number of coordinate updates of MAGIQ with the aforementioned fine tuning for $I = 6$ iterations for 16-QAM, over the whole SNR range -10 dB to 14 dB. The worst case number of loop passes is therefore $N \cdot I = 128 \cdot 6 = 768$, but the average number of loop passes is an order of magnitude lower than the worst case. Fig. 3.8 shows the GMI for 16 and 64-QAM for MAGIQ with and without the thresholding operation. We show that 64-QAM exhibits the largest gap in performance compared to no thresholding. The reduction in computational complexity comes at a price. However, even at the aggressive levels we have set in this example, there is only a 0.5 dB gap at a spectral efficiency of 5.4 bpcu, corresponding to a code rate of 0.9, which is a reasonable operating point for a coded system. However, in the range 3 bpcu to 4.5 bpcu (corresponding to code rates of 0.5 to 0.75), the gap is insignificant. For 16-QAM there is virtually no penalty across the entire SNR range. The optimized value of the threshold is 10^{-4} for 16-QAM and 10^{-6} for 64-QAM.

The ADMM algorithm, in its least complex implementation (denoted as IDE2 in [72]), has a complexity of $4NK + 3N$ multiplications for the first iteration, and another $2NK + N$ multiplications for each new iteration. The reason is that initially computed quantities can be cached and then used as memory calls in later iterations. We note that the computational burden of calculating the matrix $\boldsymbol{D}$ could be dis-

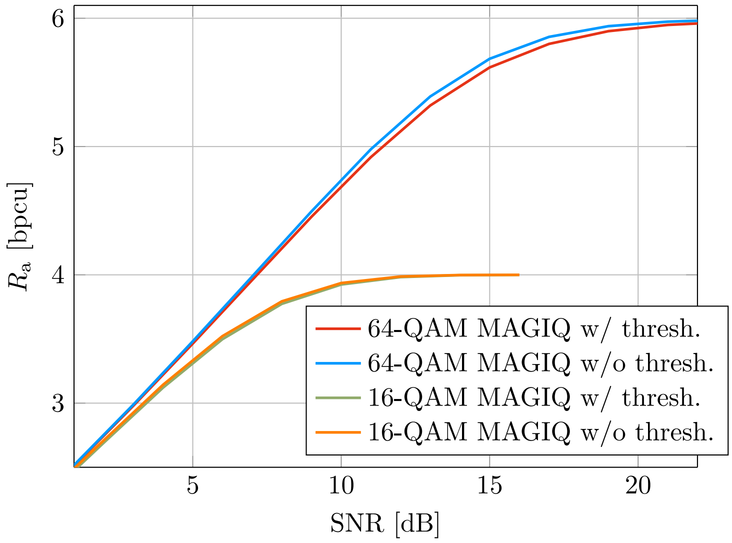


Figure 3.8.: GMI rates for $N = 128$ and $K = 16$.

counted, since it could be approximated by results in random matrix theory, as long as the statistics of $\boldsymbol{H}$ are known and the channel is stationary. For SQUID there are $NK^2 + K^3 + 3NK$ multiplications in the pre-processing phase. After pre-processing, there are $2NK + N$ multiplications per iteration. If the precoding factor α is updated for ADMM and SQUID, we consider I outer iterations that correspond to the number of times α is updated, and J inner iterations to compute $\boldsymbol{x}$ for a fixed value of α. Unlike MAGIQ that has a variable number of loop passes that depend on stopping criteria, ADMM and SQUID have fixed complexity once the number of iterations is fixed.

For frequency selective channels, we focus on the relatively simpler rules of QCM outlined in Sec. 3.5.3 and compare their performance with SQUID [42] and MSM [75] extensions for OFDM. The complexity of SQUID is thoroughly described in [42]. On the other hand, the computation complexity of MSM depends strongly on the chosen solver. In [75], the authors present a complexity analysis based on a particular instance of the simplex algorithm. We find by computer simulations that the simplex algorithm requires a very large number of iterations for convergence. In addition, the number of iterations is proportional to the number of variables and linear inequalities which grows with the size of the system (N, K, L, T). On the other hand, an interior

Table 3.1.: Computational complexity in multiplications for frequency flat channels

Algorithm	Total No. of multiplications
SQUID	$I \cdot (NK^2 + K^3 + 3NK) + I \cdot J \cdot (2NK + N)$
SQUID Num. example Fig. 3.20	1.01×10^6
ADMM	$I \cdot (4NK + 3N) + I \cdot J \cdot (2NK + N)$
ADMM Num. example Fig. 3.20	5.08×10^5
MAGIQ (worst case)	$I \cdot (3NK) + NK + J \cdot 0$
MAGIQ Num. example Fig. 3.20	10^4
MAGIQ (average, $J_{av.} = 62$) Fig. 3.7-3.8	10^4

Table 3.2.: Computational complexity per iteration for frequency selective channels

Algorithm	Complex Multiplications including Preprocessing (in red)	Iterations
SQUID	$2T \cdot (\frac{5}{3}K^3 + 3NK^2 + (6N - \frac{2}{3})K) + \mathcal{O}(8NTK + 8NT\log_2 T)$	20-50
MSM	$4KNT + \mathcal{O}(4NKT^2 + 4KT + 2NT)$	$\approx$ 8400
QCM	$KNT + 4NT\log_2 T + \mathcal{O}(NKLT + NKL\|\mathcal{X}\|)$	4-6

point solver converges much faster, but has a much higher per iteration complexity.

For QCM, by inspecting (3.60) we note that updating the current vector $\boldsymbol{x}[t]$ requires updating only L of the T terms, each a Euclidean norm. Terms of the form $\boldsymbol{u}[t]^{\mathrm{H}}\boldsymbol{u}[t]$ do not depend on $\boldsymbol{x}[t]$ and therefore are not involved in the maximization process; terms like $\|\alpha\boldsymbol{H}\boldsymbol{x}\|_2^2$ can be pre-computed and stored (with a complexity of $NKL|\mathcal{X}|$) and then reused as they do not change during one iteration; the inner product $\alpha\widetilde{\boldsymbol{u}}^{\mathrm{H}}\boldsymbol{H}\boldsymbol{x}$, however, needs to be computed for each of the L terms for each antenna update (N in total) and for each time instance, resulting in a complexity of order $\mathcal{O}(NKL)$ per time sample. To this we add the cost of initialization which is NKT plus the cost of transforming the solutions to the time domain. We discount the cost of updating the precoding factor α as all the terms involved in its computation are already available as a byproduct of the iterative process over the time samples.

Table 3.1 presents the computational complexity analysis for frequency flat channels, including numerical results shown in our examples. For frequency selective channels, the intricacy of the proposed algorithms increases and we avoid giving an exact number of multiplications but rather an order of complexity that is useful in ranking the presented algorithms. The results are summarized in Table 3.2

3.7. MSE Upper and Lower Bounds

In this section we present methods to bound and evaluate the MSE performance of the algorithms proposed so far for the optimization problem stated in (3.9). The approach taken here has been extensively used in the signal processing and communication literature. The lower bounds are based on relaxing the non-convex constraints and reformulating the cost function as a Semi-Definite Programming (SDP). This is known as a Semi-Definite Relaxation (SDR). SDR recovers a feasible solution to (3.9) by solving first the SDP and then applying a randomization and rounding procedure. Further details regarding the methods are given below as well as in Appendix B.

SDR is attractive because it belongs to the well-studied class of convex optimization problems, and can therefore be solved with polynomial complexity algorithms for a given arbitrary accuracy [89]. In some instances, it also provides theoretical guarantees on the solutions approximated by the SDR with respect to the optimal solution of the nonconvex problem and the SDP solution [90–92].

3.7.1. Bounds Based on Set Relaxation

We first derive a complex SDP relaxation based on [93]. Note that [65] also considered a real-valued SDR for the antipodal binary alphabet $\mathcal{X} = \{-1, +1\}/\sqrt{2}$. We modify the problem slightly by excluding the 0 symbol:

$$\mathcal{X} = \left\{ \sqrt{\frac{1}{N}} e^{\mathrm{j}\, 2\pi q/2^b} \text{ with } q = 0, 1, \ldots, 2^b - 1 \right\}.$$

We start from our original problem (3.9):

$$\begin{aligned} \min_{\boldsymbol{x},\alpha} \quad & \|\boldsymbol{u} - \alpha \boldsymbol{H}\boldsymbol{x}\|_2^2 + \alpha^2 K \sigma^2 \\ \text{s.t.} \quad & \boldsymbol{x} \in \mathcal{X}^N \\ & \alpha > 0 \end{aligned} \tag{3.63}$$

and proceed by defining a new variable $\boldsymbol{z} = \alpha \boldsymbol{x}$. The term $\alpha^2 K \sigma^2$ can be written as $K\sigma^2 \|\boldsymbol{z}\|_2^2$ because of the constant envelope of property of $\mathcal{X}$. In addition, we relax the discrete set membership to a continuous modulus constraint:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \|\boldsymbol{u}-\boldsymbol{H}\boldsymbol{z}\|_2^2 + K\sigma^2 \|\boldsymbol{z}\|_2^2 \\ \text{s.t.} \quad & |z_1| = |z_i|, \text{ for } i = 2, \ldots, N. \end{aligned} \tag{3.64}$$

The program (3.64) is not a homogenous quadratic problem (it contains linear terms, not only quadratic terms). However, it can be converted into a quadratic problem by an extra variable c with constant modulus $|c| = 1$. While this addition may change the optimal solution of the non-convex problem (3.64), it does not change the optimal value for the solution. This is easily seen by the unitary invariance of the squared Euclidian norm. We reformulate our cost function into a homogeneous quadratic problem of dimension $N+1$:

$$\min_{\boldsymbol{z},c} \quad \begin{bmatrix} \boldsymbol{z}^{\mathrm{H}} & c^* \end{bmatrix} \begin{bmatrix} \boldsymbol{H}^{\mathrm{H}}\boldsymbol{H} + K\sigma^2\boldsymbol{I} & -\boldsymbol{H}^{\mathrm{H}}\boldsymbol{u} \\ -\boldsymbol{u}^{\mathrm{H}}\boldsymbol{H} & \|\boldsymbol{u}\|_2^2 \end{bmatrix} \begin{bmatrix} \boldsymbol{z} \\ c \end{bmatrix} \tag{3.65}$$

with

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{H}^{\mathrm{H}}\boldsymbol{H} + K\sigma^2\boldsymbol{I} & -\boldsymbol{H}^{\mathrm{H}}\boldsymbol{u} \\ -\boldsymbol{u}^{\mathrm{H}}\boldsymbol{H} & \|\boldsymbol{u}\|_2^2 \end{bmatrix}. \tag{3.66}$$

Finally, we restate the optimization problem by introducing the matrix variable $\boldsymbol{Y} = [\boldsymbol{z} \ \ c]^{\mathrm{H}}[\boldsymbol{z} \ \ c]$. The rank of this matrix is constrained to be 1, in accordance with the rank of the optimal solution of the original problem:

$$\begin{aligned} \min_{\boldsymbol{Y}} \quad & \operatorname{Trace}(\boldsymbol{Q}\boldsymbol{Y}) \\ \text{s.t.} \quad & Y_{1,1} = Y_{i,i}, \quad \text{for } i = 2, \ldots, N \\ & Y_{N+1,N+1} = 1 \\ & \boldsymbol{Y} \succeq 0 \\ & \operatorname{rank}(\boldsymbol{Y}) = 1. \end{aligned} \tag{3.67}$$

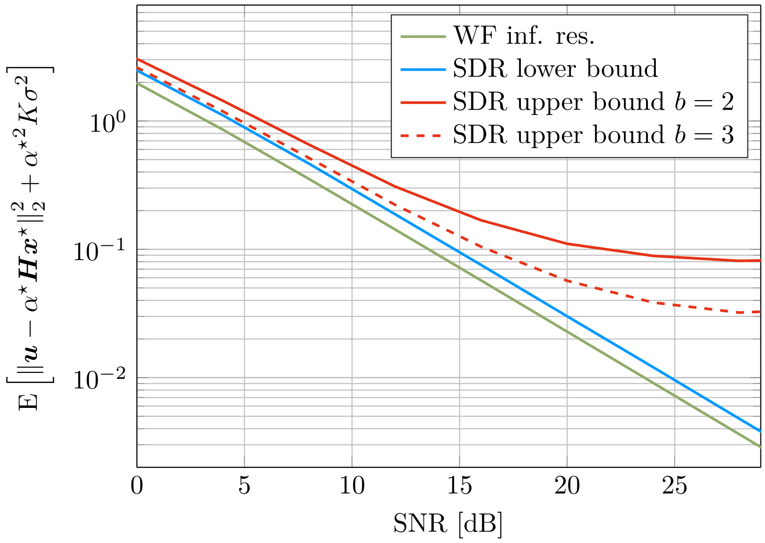


Figure 3.9.: SDR upper and lower bounds to (3.63).

Discarding the non-convex rank constraint, we obtain the SDR relaxation:

$$\begin{aligned} \min_{\boldsymbol{Y}} \quad & \mathrm{Trace}\,(\boldsymbol{QY}) \\ \text{s.t.} \quad & Y_{1,1} = Y_{i,i}, \quad \text{for } i = 2, \dots, N \\ & Y_{N+1,N+1} = 1 \\ & \boldsymbol{Y} \succeq 0. \end{aligned} \tag{3.68}$$

The solution of the minimization above is a lower bound to (3.63) by virtue of enlarging the constraint set and properties of convex optimization problems [89]. If the optimal $\boldsymbol{Y}^\star$ is of rank one after solving (3.68), then the solution to the relaxed problem (3.64) can be directly recovered from the eigenvector corresponding to the single non-zero eigenvalue. Furthermore, to obtain a feasible solution for (3.63) we project component-wise onto the feasible set.

If the solution is not rank one, then feasible solutions and a selection of the best is obtained by the randomization algorithm presented in Table I in [93]. This method gives an upper bound to (3.63). We add that SQUID, ADMM, MAGIQ and QCM all obtain feasible solutions for (3.63) and therefore upper bounds to the cost function. Fig. 3.9 shows the lower bound to (3.63) obtained from (3.68). Note that in the low

SNR regime the bounds are tight, while at high SNR (for higher rates) the bounds become loose and uninformative. The lower bound is unaffected by the increasing cardinality of the transmit set $\mathcal{X}$ as it assumes a relaxed transmit set on a disc centered at the origin. Note also that there is a constant gap of approximately 1.5dB between the SDP solution and the WF, suggesting that a constant envelope signaling method would be fundamentally bounded away from an infinite resolution WF.

3.7.2. Improved Bounds

We tighten the lower bound by an approach that extends the methods presented in [94–97]. Consider the cost function:

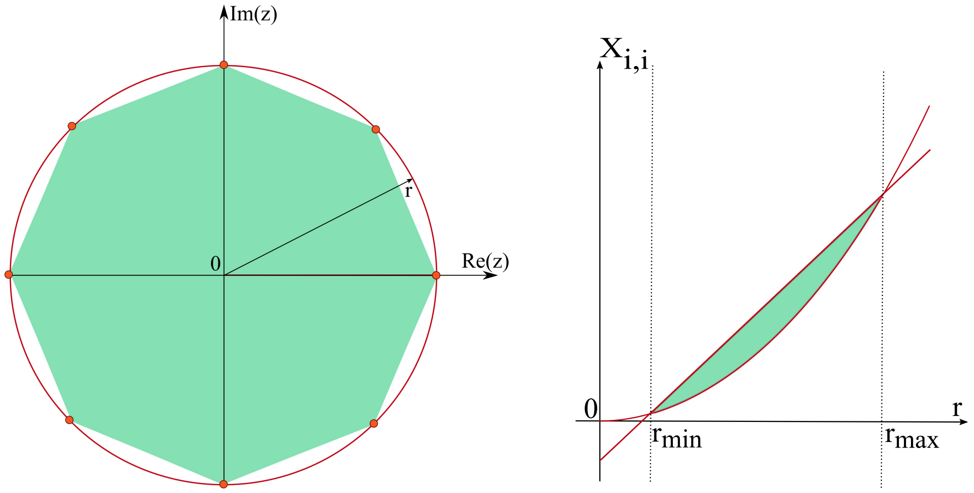
$$\|\boldsymbol{u}-\boldsymbol{H}\boldsymbol{z}\|_2^2 + K\sigma^2\|\boldsymbol{z}\|_2^2 = \boldsymbol{z}^{\mathrm{H}}(\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}+K\sigma^2\boldsymbol{I})\boldsymbol{z} - 2\,\mathrm{Re}\{\boldsymbol{z}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{u}\} + \|\boldsymbol{u}\|_2^2 \tag{3.69}$$

The last term does not depend on $\boldsymbol{z}$ and therefore does not impact the optimization. We introduce the variable $\boldsymbol{X}=\boldsymbol{z}\boldsymbol{z}^{\mathrm{H}}$ and perform the SDP relaxation $\boldsymbol{X}\succeq\boldsymbol{z}\boldsymbol{z}^{\mathrm{H}}$. With these modifications we reformulate (3.64) as

$$\begin{aligned}\min_{\boldsymbol{X},\boldsymbol{z}}\quad & \mathrm{Trace}\left((\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}+K\sigma^2\boldsymbol{I})\boldsymbol{X}\right) - 2\,\mathrm{Re}\{\boldsymbol{z}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{u}\} + \|\boldsymbol{u}\|_2^2\\ \text{s.t.}\quad & X_{1,1}=X_{i,i},\ \ \text{for } i=2,\dots,N\\ & \boldsymbol{X}\succeq\boldsymbol{z}\boldsymbol{z}^{\mathrm{H}}.\end{aligned}\tag{3.70}$$

We have now a SDP in two variables, $\boldsymbol{X}$ and $\boldsymbol{z}$, connected through the positive semidefinite constraint.

However, (3.70) does not preserve information about the discrete nature of the original set $\mathcal{X}$. We introduce linear constraints on $\boldsymbol{z}$ that are consistent with a tight convex relaxation of $\mathcal{X}$. Let $\mathcal{A}_{const}$ be the convex hull of the discrete set $\alpha\mathcal{X}$, e.g., $\mathcal{A}_{const}$ is illustrated in Fig. 3.10(a) for $b=3$. Next, since each element in $\alpha\mathcal{X}$ has a constant modulus that is part of our optimization, we introduce the variable $r=|z_i|$. The constraints $X_{1,1}=X_{i,i},\ \ \text{for } i=2,\dots,N$ and $\boldsymbol{X}\succeq\boldsymbol{z}\boldsymbol{z}^{\mathrm{H}}$ can be restated as $X_{i,i}\geq r^2,\ \ \text{for } i=2,\dots,N$. Moreover, this constraint can be tightened by noticing that r can be upper and lower bounded as $\mathrm{r}_{min}\leq r\leq\mathrm{r}_{max}$. A trivial lower bound is 0. For the upper bound, at high SNR a simple derivation gives $r\leq\|u\|_2^2/\sqrt{N}$. The



(a) The constraint set $\mathcal{A}_{const}$ for $b = 3$

(b) The constraint set $\mathcal{G}_{rX}$

Figure 3.10.: Convex envelopes by linear inequalities.

set that summarizes these constraints is illustrated in Fig. 3.10(b) and is given by:

$$\mathcal{G}_{rX} = \Big\{(X_{i,i}, r) \mid X_{i,i} \geq r^2 \wedge X_{i,i} - (\mathrm{r}_{min} + \mathrm{r}_{max})r + \mathrm{r}_{min}\mathrm{r}_{max} \leq 0\Big\}. \tag{3.71}$$

We are now ready to give the final tightened complex SDR formulation (TCSDR):

$$\begin{aligned}
\min_{\boldsymbol{X},\boldsymbol{z},r} \quad & \mathrm{Trace}\Big((\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H} + K\sigma^2\boldsymbol{I})\boldsymbol{X}\Big) - 2\,\mathrm{Re}\{\boldsymbol{z}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{u}\} + \|\boldsymbol{u}\|_2^2\,. \\
\text{s.t.} \quad & (r, z_i) \in \mathcal{A}_{const}, \quad \text{for } i = 1, \ldots, N \\
& (X_{i,i}, r) \in \mathcal{G}_{rX}, \quad \text{for } i = 1, \ldots, N \\
& \mathrm{r}_{min}^2 \leq X_{i,i} \leq \mathrm{r}_{max}^2 \\
& \boldsymbol{X} \succeq \boldsymbol{z}\boldsymbol{z}^{\mathrm{H}}.
\end{aligned} \tag{3.72}$$

The TCSDR (3.72) solution is used in a Gaussian randomized procedure, given by Algorithm 5, to obtain feasible solutions and upper bounds.

We also tighten the upper bound by starting from the already proposed MAGIQ algorithm and extend it to cover a larger search space. Since the decision at the first layer is based on a best first strategy, and any of the unselected runner-up coordinates can be very close to the winning candidate, we proceed with a subset (a list) of the

Algorithm 5: Semidefinite Relaxation with Gaussian Randomization

```
Solve the TCSDR program, obtain optimal solutions $\boldsymbol{X}^\star$, $\boldsymbol{z}^\star$
Compute Cholesky factorization of $\boldsymbol{M}\boldsymbol{M}^{\mathrm{T}} = \boldsymbol{X}^\star - \boldsymbol{z}^\star\boldsymbol{z}^{\star\mathrm{H}}$
for $samples = 1 : \mathrm{R}$ do
    Generate a random sample $\boldsymbol{r} = \boldsymbol{z}^\star + \boldsymbol{M}\boldsymbol{w}$ where $\boldsymbol{w} \sim \mathcal{CN}(0, I)$
    Project onto feasible set $\boldsymbol{r}_\Pi = \Pi_{\mathcal{X}}\{\boldsymbol{r}\}$
    If $F(\boldsymbol{r}_\Pi, r) < \mathrm{F}_{min}$ then $\boldsymbol{x} = \boldsymbol{r}_\Pi$, $\alpha = r$, $\mathrm{F}_{min} = F(\boldsymbol{x}, \alpha)$
end
Output: $\boldsymbol{x}$,$\alpha$,$\mathrm{F}_{min}$
```

unselected coordinates at each iteration and repeat the MAGIQ procedure for the selected list. The method is presented in Algorithm 6 and is called L-MAGIQ.

Finally, to validate the hypothesis that improved performance at high SNR demands considerably more computational complexity for a fixed bit-width, we examine a quasi-Maximum Likelihood (ML) algorithm. Sphere Precoding (SP) is known as an efficient implementation of ML algorithms for Closest Vector Problem (CVP) and Shortest Vector Problem (SVP) [98]. The core idea of SP is to constrain the search over a radius that can be adapted as the algorithm advances. The SP algorithm has an average polynomial complexity for finding the ML solution for smaller problems at high SNR, but still with a worst case exponential complexity, similar to exhaustive search [99]. We reformulate the precoding problem so that we can apply the formalism of SP:

$$
\begin{aligned}
F(\boldsymbol{x}, \alpha) &= \|\boldsymbol{u} - \alpha\boldsymbol{H}\boldsymbol{x}\|_2^2 + \alpha^2 K\sigma^2 && (3.73)\\
&= \|\boldsymbol{u} - \alpha\boldsymbol{H}\boldsymbol{x}\|_2^2 + \alpha^2 K\sigma^2 \frac{\|x\|^2}{P} && (3.74)\\
&= \left\|\hat{\boldsymbol{u}} - \alpha\hat{\boldsymbol{H}}\boldsymbol{x}\right\|_2^2 && (3.75)
\end{aligned}
$$

where $\hat{\boldsymbol{u}} = [\boldsymbol{u}^{\mathrm{T}} \quad \mathbf{0}^{\mathrm{T}}]^{\mathrm{T}}$ and $\hat{\boldsymbol{H}} = [\boldsymbol{H}^{\mathrm{T}} \quad \sqrt{K\sigma^2/P}\mathbf{I}^{\mathrm{T}}]^{\mathrm{T}}$. With the QR factorization of $\hat{\boldsymbol{H}}$ we can further write:

$$
\left\|\hat{\boldsymbol{u}} - \alpha\hat{\boldsymbol{H}}\boldsymbol{x}\right\|_2^2 = \left\|\boldsymbol{Q}^{\mathrm{H}}\hat{\boldsymbol{u}} - \alpha\boldsymbol{R}\boldsymbol{x}\right\|_2^2 = \|\tilde{\boldsymbol{u}} - \alpha\boldsymbol{R}\boldsymbol{x}\|_2^2 \tag{3.76}
$$

Algorithm 6: L-MAGIQ

```
Inputs: $\boldsymbol{u}$, $\boldsymbol{H}$, $\mathcal{S} = \{1, \dots, N\}$, $err_{\min}$
Initialize: $\boldsymbol{x} = \boldsymbol{x}_{init}$, $\alpha = \alpha_{init}$, $\mathcal{L} = \emptyset$, $\mathcal{E} = \emptyset$, $\mathcal{B} = \emptyset$, $\mathcal{T} = \emptyset$
for $j = 1 : L$ do
    $\mathcal{S} = \{1, \dots, N\}$
    for $i = 1 : I$ do
        $n = 1$
        $err_0 = \|\boldsymbol{u}\|^2$
        while $(err_n > err_{min}) \vee (err_n < err_{n-1}) \vee (n \leq N)$ do
            if $n = 1 \wedge i = 1$ then
                $(x^\star_{n^\star}, n^\star) = \text{argmin}_{x_n \in \mathcal{X}, n \in \mathcal{S} \setminus \mathcal{L}} F(\boldsymbol{x}, \alpha)$
                $(x_1, \dots, x_n, \dots, x_N)^{\mathrm{T}} = (x_1, \dots, x^\star_{n^\star}, \dots, x_N)^{\mathrm{T}}$
                $\mathcal{L} \leftarrow \mathcal{L} \cup \{n^\star\}$
            else
                $(x^\star_{n^\star}, n^\star) = \text{argmin}_{x_n \in \mathcal{X}, n \in \mathcal{S}} F(\boldsymbol{x}, \alpha)$
                $(x_1, \dots, x_n, \dots, x_N)^{\mathrm{T}} = (x_1, \dots, x^\star_{n^\star}, \dots, x_N)^{\mathrm{T}}$
            end
            $\mathcal{S} \leftarrow \mathcal{S} \setminus \{n^\star\}$
            $n \leftarrow n + 1$
            $err_n = \|\boldsymbol{u} - \alpha \boldsymbol{H}\boldsymbol{x}\|_2^2 + \alpha^2 K \sigma^2$
        end
        $\alpha = \frac{\mathrm{Re}(\boldsymbol{u}^{\mathrm{H}} \boldsymbol{H}\boldsymbol{x})}{\|\boldsymbol{H}\boldsymbol{x}\|_2^2 + K\sigma^2}$
        $\mathcal{B} = \mathcal{B} \cup \{\boldsymbol{x}\}$
        $\mathcal{E} = \mathcal{E} \cup \{err_N\}$
        $\mathcal{T} = \mathcal{T} \cup \{\alpha\}$
    end
end
Output $\mathcal{B}, \mathcal{E}, \mathcal{T}$
```

The SP problem is finally formulated as:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \|\tilde{\boldsymbol{u}} - \alpha \boldsymbol{R}\boldsymbol{x}\|_2^2 \\ \text{s.t.} \quad & \boldsymbol{x} \in \mathcal{X}^N. \end{aligned} \tag{3.77}$$

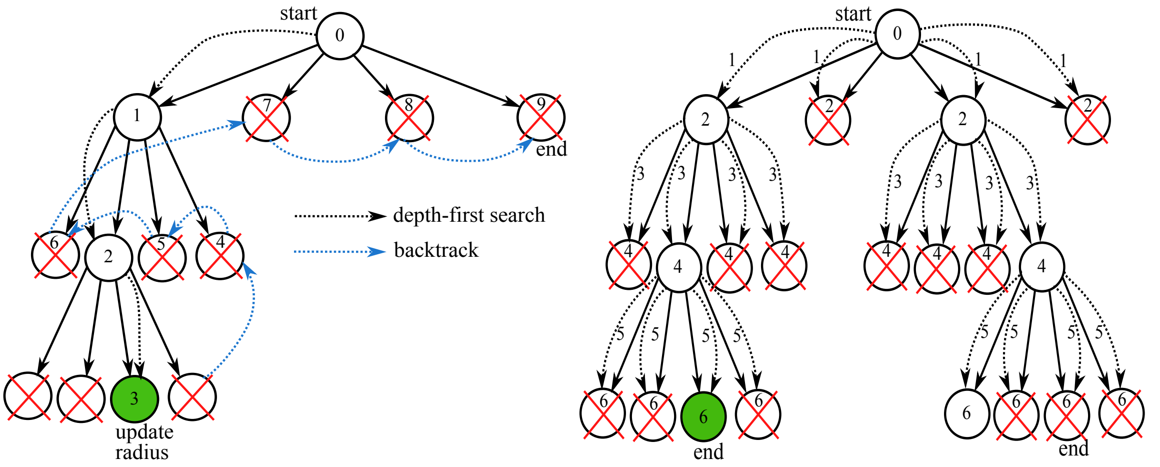
We approach the solution with the K-Best [100] tree search algorithm. The SP traverses the tree depth-first (from the root to a leaf) and then backtracks to upper levels of the tree after updating the search radius each time it encounters a leaf. This allows SP to prune a significant number of nodes from the tree and thus reduce the search complexity. However, the number of leaves grows exponentially with the depth of the tree (number of antennas) and although we have tested SP for our massive MIMO scenario, the run-time of the algorithm is prohibitive and could not be evaluated.

Fig. 3.11 shows both the operation of SP and K-Best. K-Best processes nodes in the tree in parallel one level at a time. After computing the partial branch metric at the current level it selects K nodes with minimum metric that are then considered for expansion for the next level. This process continues until the leaves of the tree, where we are left with K candidates and the declared solution is the node with the minimum accumulated metric. This procedure reveals that K-Best operates at each step based on only partial knowledge (available at the current level) and could miss the ML solution. The likelihood of the true ML solution being in the candidate set grows with K and therefore K-Best with a sufficiently large K operates very close to SP (when K is equal to the number of leaves in the tree, K-Best is identical to exhaustive search). We obtain our numerical results with a value of K that allows for an acceptable runtime of the algorithm.

Improving the K-Best decisions made at the levels closest to the root significantly impacts the quality of the solution. We take a twofold approach to alleviate this issue. First, we retain all nodes at the first levels of the tree in the candidate list (we do this for the first 4 levels). In addition, we use a sorted QR decomposition, that orders the diagonal values of R in an ascending order. This results in partial branch metrics at the early levels of the tree being a substantial portion of the final metric, i.e., decisions taken at early levels are less likely to be changed later in the tree search. These two procedures significantly improve K-Best detectors for MIMO receivers [101].

Note that the precoding factor α is not part of the optimization problem (3.77). The optimization with respect to α follows the procedure outlined for MAGIQ, alternating between updates of $\boldsymbol{x}$ and α, until convergence is achieved or a predefined number of iterations is performed. Thus, even if SP provides the optimal solution to (3.77), the alternating procedure does not necessarily provide the optimal solution of (3.9).

Fig. 3.12 and Fig. 3.13 show the improved upper (L-MAGIQ with $L = 64$ and



(a) SP depth-first tree search with radius update

(b) K-Best breath-first tree search

Figure 3.11.: Optimal (a) and quasi-optimal (b) tree search algorithms for solving the quantized precoding problem. The tree traversal schedule is shown with dashed arrows, the order of exploration is marked with numbers and the pruned branches from the search are marked with a red 'X'. The leaf corresponding to the final solution is marked with green.

K-Best with $K = 4096$ and 6 iterations) and lower bounds (TCSDR) as well as the results returned by MAGIQ and ADMM. Note that, for an alphabet $\mathcal{X}$ supported on 4 mass points ($b = 2$), the new lower bound offers a 4dB improvement on the SDR relaxation (3.68) at $MSE = 10^{-2}$. For the upper bound, L-MAGIQ offers a 27% improvement over MAGIQ, while K-Best provides a significant 44% decrease in MSE at an SNR of 24 dB. For $b = 3$ we observe that the upper bounds are considerably closer to the SDR lower bound over a large SNR range and the tighter lower and upper bounds (TCSDR and K-Best) offer an improvement of 68% at an SNR of 24 dB. This is important from an algorithmic point of view because the upper bounds are always achievable and inform us that more sophisticated algorithms or increased resolution are needed for higher rates. Interestingly, all the bounds are tight with the SDR lower bound up until $10 - 12$dB showing that all the proposed algorithms are nearly optimal for the chosen metric at low and medium rates. These results are consistent with the lower bounds to information rates shown below in Sec. 3.10.

Fig. 3.13 also shows the upper bound obtained with the infinite resolution ($b = \infty$)

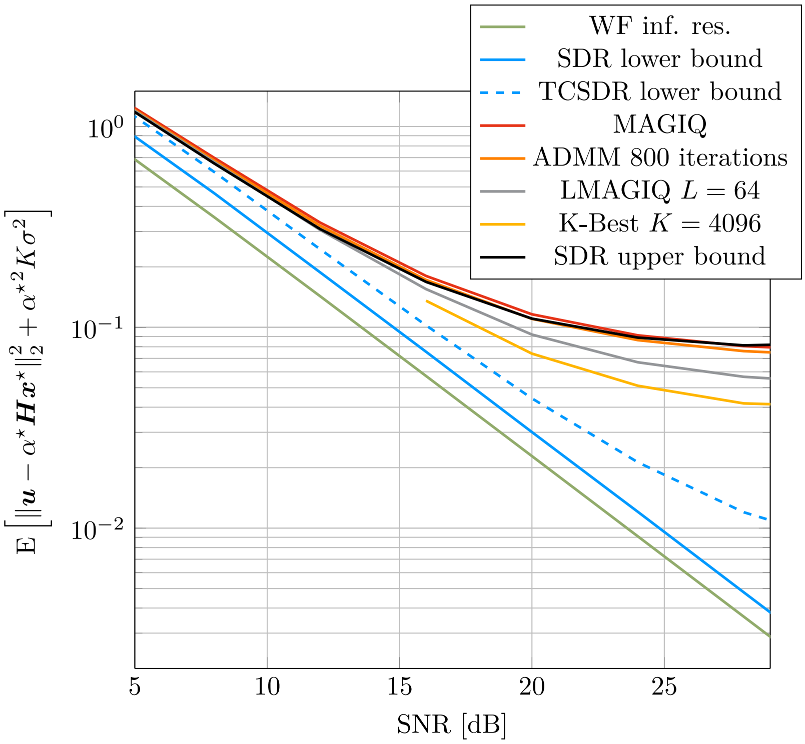


Figure 3.12.: Tightened upper and lower bounds for $b = 2$.

precoder from [102]. Observe that the gap to the SDR lower bound is negligible at almost all operational SNR levels, demonstrating that a simple projected coordinate descent algorithm with a constant modulus constraint achieves the SDR lower bound. This provides a quantitative evaluation of the constant gap to WF for constant envelope signaling.

3.8. Sensitivity to Channel Uncertainty

All results presented so far assume that the transmitter has perfect CSI, i.e., the channel matrix $\boldsymbol{H}$ that the precoder uses is the same as the channel realization. However, in practice the base station has access to imperfect channel knowledge due to noise, quantization, channel calibration errors, etc. We do not try to model these effects exactly. Instead, we adopt a common approach for evaluating robustness that

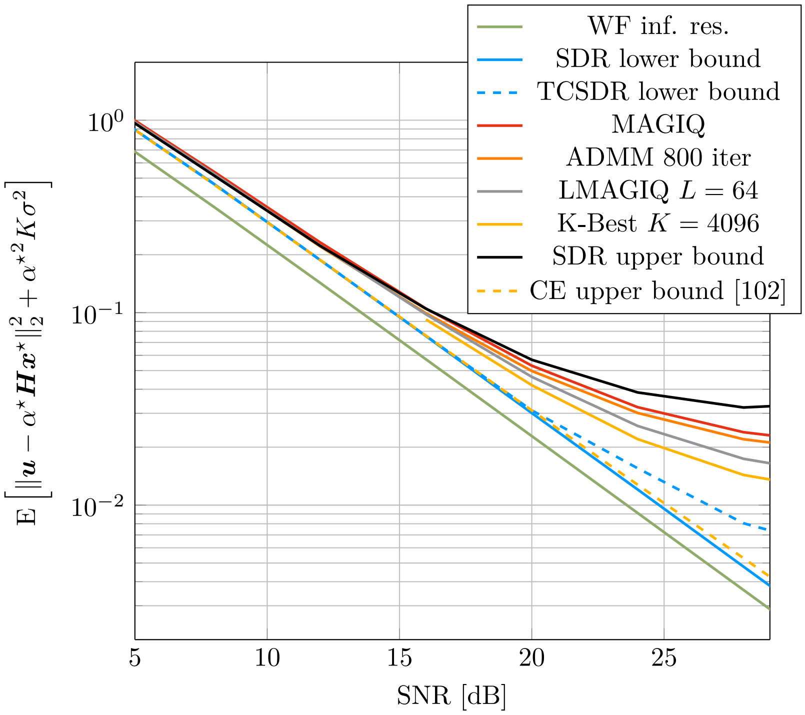


Figure 3.13.: Tightened upper and lower bounds for $b = 3$.

provides the precoder with a noisy channel matrix $\hat{\boldsymbol{H}}$ that satisfies

$$\boldsymbol{H} = \sqrt{1-\varepsilon}\hat{\boldsymbol{H}} + \sqrt{\varepsilon}\boldsymbol{Z} \tag{3.78}$$

where $\varepsilon \in [0, 1]$ and $\boldsymbol{Z}$ is a $K \times N$ matrix of mutually independent, zero-mean, variance σ_h^2 Gaussian entries. Taking expectations over $\boldsymbol{Z}$, we thus have

$$\mathrm{E}\left[\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}\right] = (1-\varepsilon)\hat{\boldsymbol{H}}^{\mathrm{H}}\hat{\boldsymbol{H}} + \varepsilon\boldsymbol{C}_z \tag{3.79}$$

where $\boldsymbol{C}_z$ is an $N \times N$ covariance matrix. We refer to ε as the channel-estimation power: $\varepsilon = 0$ corresponds to perfect CSI and $\varepsilon = 1$ corresponds to no CSI.

We first consider the case where the precoder considers the estimated channel matrix $\hat{\boldsymbol{H}}$ to be the actual channel realization $\boldsymbol{H}$. Fig. 3.14 shows an example of the degradation of mutual information rates for all non-linear precoders presented in this report and benchmarks them against the performance of an infinite resolution ZF. The

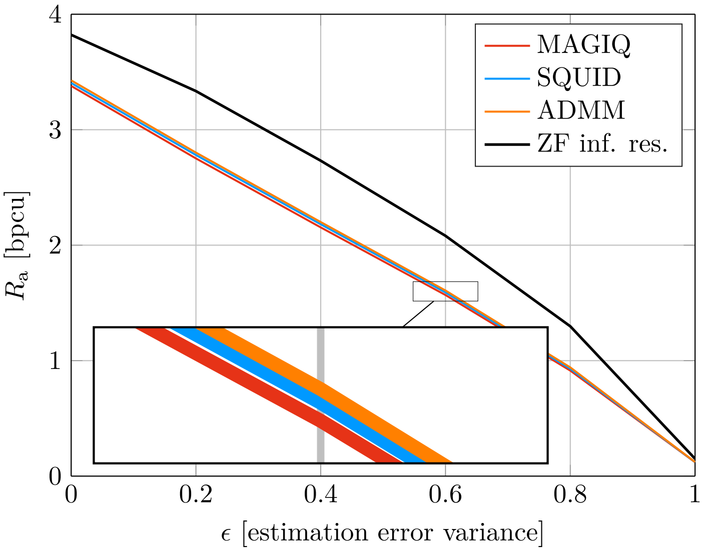


Figure 3.14.: GMI rates for 16-QAM, $N = 128$, and $K = 16$ vs ε

simulation is done for the particular case of $\boldsymbol{C}_z = \sigma_h^2 \boldsymbol{I}$. There is a constant decay of rates with the variance of the disturbance added to the perfect channel knowledge, which shows that the performance degradation of the precoders is both continuous and graceful.

We now derive a robust MAGIQ (QCM) and show that, within the framework of MMSE precoding, the effect of the channel estimation error can be efficiently incorporated into the cost function with negligible computational penalty. We compute:

$$\begin{aligned} MSE &= \mathrm{E}\left[\|\boldsymbol{u} - \alpha \boldsymbol{H}\boldsymbol{x}\|_2^2 + \alpha^2 K \sigma^2\right] \\ &= \mathrm{E}\left[(\boldsymbol{u} - \alpha \boldsymbol{H}\boldsymbol{x})^{\mathrm{H}} (\boldsymbol{u} - \alpha \boldsymbol{H}\boldsymbol{x}) + \alpha^2 K \sigma^2\right] \\ &= \mathrm{E}\left[\left(\boldsymbol{u} - \alpha(\sqrt{1-\varepsilon}\hat{\boldsymbol{H}} + \sqrt{\varepsilon}\boldsymbol{Z})\boldsymbol{x}\right)^{\mathrm{H}} \left(\boldsymbol{u} - \alpha(\sqrt{1-\varepsilon}\hat{\boldsymbol{H}} + \sqrt{\varepsilon}\boldsymbol{Z})\boldsymbol{x}\right) + \alpha^2 K \sigma^2\right] \\ &= \boldsymbol{u}^{\mathrm{H}}\boldsymbol{u} - \sqrt{1-\varepsilon}\alpha \boldsymbol{u}^{\mathrm{H}}\hat{\boldsymbol{H}}\boldsymbol{x} - \sqrt{1-\varepsilon}\alpha \boldsymbol{x}^{\mathrm{H}}\hat{\boldsymbol{H}}^{\mathrm{H}}\boldsymbol{u} + \alpha^2(1-\varepsilon)\boldsymbol{x}^{\mathrm{H}}\hat{\boldsymbol{H}}^{\mathrm{H}}\hat{\boldsymbol{H}}\boldsymbol{x} \\ &\quad + \alpha^2\varepsilon \boldsymbol{x}^{\mathrm{H}}\boldsymbol{C}_z\boldsymbol{x} + \alpha^2 K \sigma^2. \end{aligned} \tag{3.80}$$

The robust MAGIQ precoder cost function has an extra quadratic term $\alpha^2\varepsilon\boldsymbol{x}^{\mathrm{H}}\boldsymbol{C}_z\boldsymbol{x}$ as compared to the non-robust MAGIQ (3.9). This term acts as a regularizer for the cost function and weighs the beamforming directions with the channel error covariance,

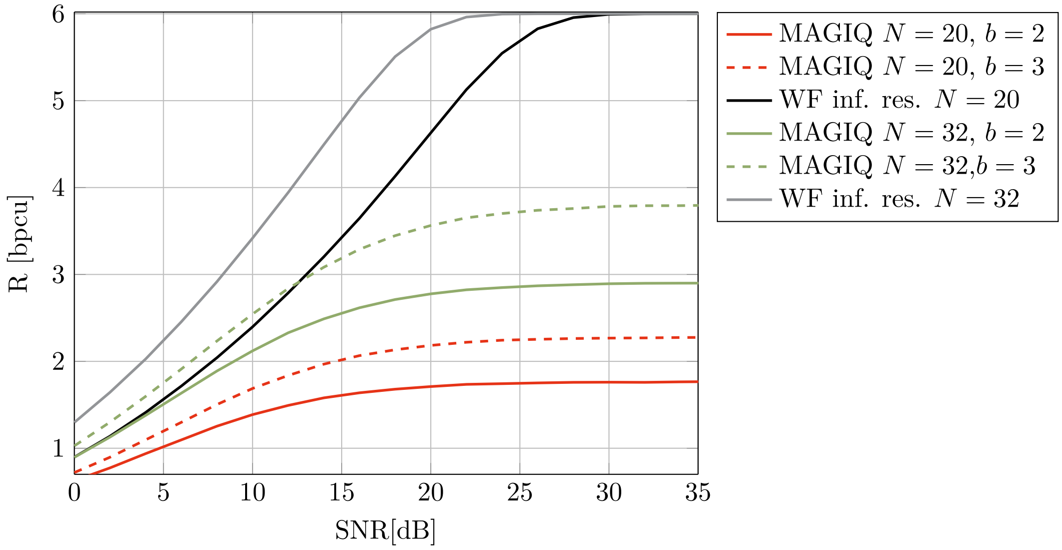


Figure 3.15.: GMI rates for 16-QAM, $K = 16$.

i.e., it modifies the directions where the norm of the channel error is large, and favours those directions with smaller norms. This is similar to results obtained in the literature for robust linear precoding [103]. Observe from (3.80) that the extra computational complexity of robust MAGIQ resides in the computation of $\boldsymbol{C}_z$ and $(1 - \varepsilon)\hat{\boldsymbol{H}}^{\mathrm{H}}\hat{\boldsymbol{H}}$.

3.9. Mixed Resolution RF Chains

Fig. 3.15 shows one important limitation of the proposed architecture and precoder. For a high system load, either because of a small number of antennas at the base station or a large number of users, the achievable rates are significantly reduced as a result of multiuser interference. As can be observed, MAGIQ with $b = 2$ and $b = 3$ and with $N = 20$ and $N = 32$ antennas performs far (at least 2.2 bits) from the WF at high SNR.

An interesting option to introduce low resolution RF chains may be low-cost plug-in upgrades of legacy infrastructure, e.g., an Long Term Evolution (LTE) evolved Node B (eNB) with high resolution RF chains upgraded to 64 additional antennas with low resolution RF chains. We thus consider precoding for antennas with both low resolution RF chains, like the ones proposed in Sec. 3.2, and high resolution

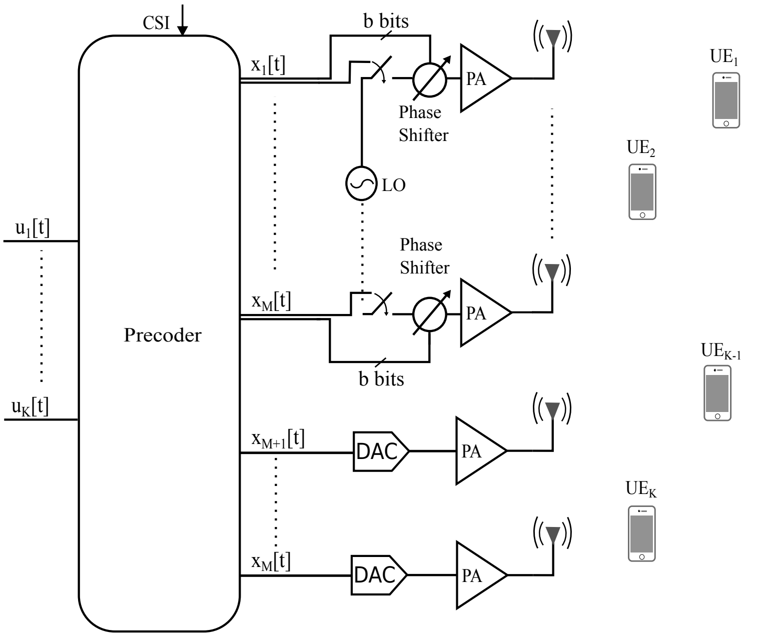


Figure 3.16.: Multi-user MIMO downlink with mixed resolution RF chains.

RF chains, e.g., 16 to 20 bits per dimension. For simplicity, we assume that each antenna is assigned a fixed RF chain, i.e., the DAC/phase shifter resolutions cannot be dynamically adjusted (such options would likely improve performance, but they are outside the scope of this work). A typical downlink system is presented in Fig. 3.16.

We denote high resolution RF chains with the acronym Full RF (FRF) and low resolution chains with Quantized RF (QRF). The precoding vector can be represented as a concatenation of two sub-vectors in the manner $\boldsymbol{x} = [\boldsymbol{x}_{QRF},\ \boldsymbol{x}_{FRF}]$. The channel matrix $\boldsymbol{H} = [\boldsymbol{H}_{QRF},\ \boldsymbol{H}_{FRF}]$ can be similarly split. We consider a single cell with $N - M$ FRF transmit antennas, M QRF transmit antennas, and K users. The modulation set for each QRF antenna is defined as:

$$\tilde{\mathcal{X}} = \{0\} \cup \left\{ \sqrt{\frac{P(1-\theta)}{M}}\, \mathrm{e}^{\mathrm{j}\, 2\pi q/2^b}; q = 1, 2, \ldots, 2^b \right\}. \tag{3.81}$$

For FRF, we consider an average power constraint $\mathrm{E}\left[\|\boldsymbol{x}_{FRF}\|_2^2\right] \leq P\theta$.

The precoding problem for flat-fading channels is based on the MSE criteria

$$\begin{aligned} \min_{\mathbf{x},\alpha,\theta} \quad & \|\boldsymbol{u} - \alpha \boldsymbol{H}_{QRF}\boldsymbol{x}_{QRF} - \alpha \boldsymbol{H}_{FRF}\boldsymbol{x}_{FRF}\|_2^2 + \alpha^2 K\sigma^2 \\ \text{subject to} \quad & \boldsymbol{x}_{QRF} \in \tilde{\mathcal{X}}^M,\ \mathrm{E}\left[\|\boldsymbol{x}_{FRF}\|_2^2\right] \leq P\theta \\ & \alpha > 0. \end{aligned} \tag{3.82}$$

Apart from the precoding vector $\boldsymbol{x}$ and the precoding factor α, we have an additional power allocation variable θ which determines what percent of the total available power is divided between the FRF and QRF sub-arrays. Because of the discrete constraints on the QRF, the problem (3.82) is a mixed-integer program and therefore also NP-hard.

The precoding problem must be solved jointly and we propose several iterative optimization algorithms. These algorithms require CSI only about individual links between the corresponding antennas and users. We propose two iterative optimization schedules:

- ▷ Minimize the cost function by updating the FRF precoding vector for each greedy choice of a single QRF antenna (FRF-MAGIQ);
- ▷ Minimize the cost function alternately between the MAGIQ solution and FRF solution (Alt-MAGIQ);
- ▷ θ is optimized offline with a line search and with the ergodic achievable rate as a cost function (rater than the MSE cost function used for precoding).

The optimization over θ could be done instantaneously for each transmitted symbol, which gives an upper bound on the performance. However, to be consistent with our CSI assumptions, we will optimize the power allocation offline and apply it only once. Fig. 3.17 shows how the average per user GMI depends on θ at a given SNR. Interestingly, the rate variation is slow and smooth, and is maximized by a ratio close to $1 - M/N$.

The iterative schedule lets us solve simplified problems with reduced computational complexity. With a QRF solution $\boldsymbol{x}^{\star}_{QRF}$, the precoding problem simplifies to

$$\|\tilde{\boldsymbol{u}} - \alpha \boldsymbol{H}_{FRF}\boldsymbol{x}_{FRF}\|_2^2 + \alpha^2 K\sigma^2$$

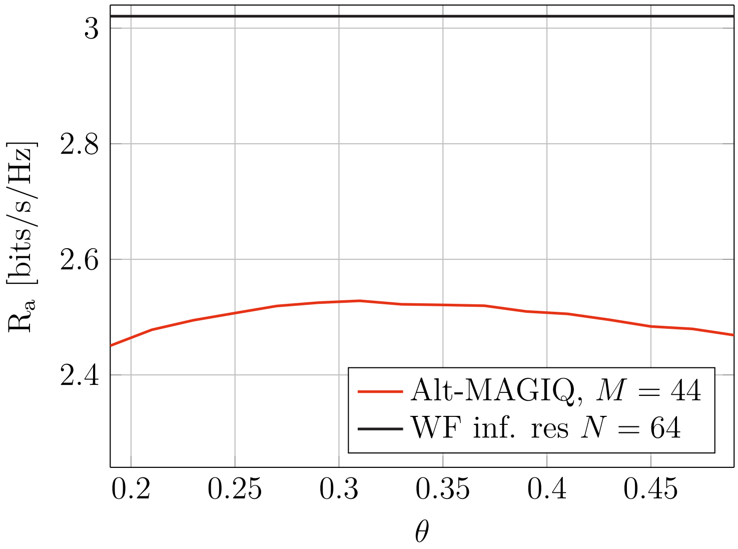


Figure 3.17.: GMI rates for 64-QAM vs. θ for $N = 64$, $M = 44$, $K = 16$, $b = 2$, $SNR = 4dB$

$$\text{where } \tilde{\boldsymbol{u}} = \boldsymbol{u} - \alpha \boldsymbol{H}_{QRF} \boldsymbol{x}^{\star}_{QRF}.$$

We further relax the cost to an obtain the well known MMSE:

$$\begin{aligned} \min_{\boldsymbol{x}_{FRF}, \alpha} \quad & \mathrm{E}\left[\|\tilde{\boldsymbol{u}} - \alpha \boldsymbol{H}_{FRF} \boldsymbol{x}_{FRF}\|_2^2 \right] + \alpha^2 K \sigma^2 \\ \text{subject to} \quad & \mathrm{E}\left[\|\boldsymbol{x}_{FRF}\|_2^2 \right] \leq P_\theta \\ & \alpha > 0. \end{aligned} \tag{3.83}$$

The solution to (3.83) is the WF presented in Sec. 3.3.2. and denoted here for FRF as $\boldsymbol{x}^{\star}_{FRF}$, the remaining problem can be solved approximately by MAGIQ. This is the basis of our algorithms as outlined in Alg. 7 and Alg. 8

We emphasize here that the main reason for proposing these solutions is to reuse MAGIQ with existing linear precoding techniques. The mixed RF problem is akin to a quadratic mixed-integer program and one could design more powerful algorithms based on state-of-the-art branch-and-bound or branch-and-cut methods [104] that can find optimal solutions to such problems. However, for our specific application, complexity and run-time are of utmost importance and we rely on sub-optimal techniques

Algorithm 7: FRF-MAGIQ for frequency-flat channels

```
Inputs: $\boldsymbol{u}$, $\boldsymbol{H}$, $\mathcal{S} = \{1, \ldots, N\}$
Initialize: $\boldsymbol{x} = \boldsymbol{x}_{init}$, $\alpha = \alpha_{init}$
for $i = 1 : I$ do
    while error decreasing do
        for $n \in \mathbb{S}$ do
            for $x_n \in \mathcal{X}$ do
                $\boldsymbol{x}_{FRF}(x_n) = \boldsymbol{H}_{FRF}^{\mathrm{H}} \left(\boldsymbol{H}_{FRF}\boldsymbol{H}_{FRF}^{\mathrm{H}} + \gamma \boldsymbol{I}\right)^{-1} (\boldsymbol{u} - \alpha^{(i)} \boldsymbol{H}_{QRF} \boldsymbol{x}_{QRF})$
            end
            $(x_n^{\star}, \boldsymbol{x}_{FRF}^{\star}) = \arg\min_{x_n, \boldsymbol{x}_{FRF}(x_n)} MSE(\boldsymbol{x}, \alpha)$
        end
        $(x_1^{(i)}, \ldots, x_n^{(i)}, \ldots, x_{N-M}^{(i)})^{\mathrm{T}} = (x_1^{(i)}, \ldots, x_{n^{\star}}^{\star}, \ldots, x_{N-M}^{(i)})^{\mathrm{T}}$
        $\boldsymbol{x}_{FRF}^{(i)} = \boldsymbol{x}_{FRF}^{\star}$
        $\mathbb{S} \leftarrow \mathbb{S} \setminus \{n^{\star}\}$
    end
    $\boldsymbol{x}^{(i)} = [\boldsymbol{x}_{QRF}^{(i)},\ \boldsymbol{x}_{FRF}^{(i)}]$
    $\alpha^{(i)} = \frac{\mathrm{Real}\{\mathbf{u}^{\mathrm{H}}\mathbf{H}\mathbf{x}^{(i)}\}}{||\mathbf{H}\mathbf{x}^{(i)}||_2^2 + K\sigma^2}$
end
Output $\boldsymbol{x}^{(I)}, \alpha^{(I)}$
```

that offer a solid trade-off between the design constraints and solution accuracy.

3.10. Numerical Results

We begin the numerical study with the frequency-flat case. We compare the performance of different precoding schemes by means of their GMI for massive MIMO with $K = 16$ UEs and $N = 128$ antennas at the base station. If not stated otherwise, all results are reported for Scenario 1 presented in Sec. 3.2.3.

3.10.1. Quantized Linear Precoding

In Fig. 3.18 and Fig. 3.19 we compare the lower bounds and signaling alphabets proposed in this work and in [65] for QLPs. The performance of the transmit WF with infinite resolution (up to machine precision) signaling with both Gaussian distributed

Algorithm 8: Alt-MAGIQ for frequency-flat channels

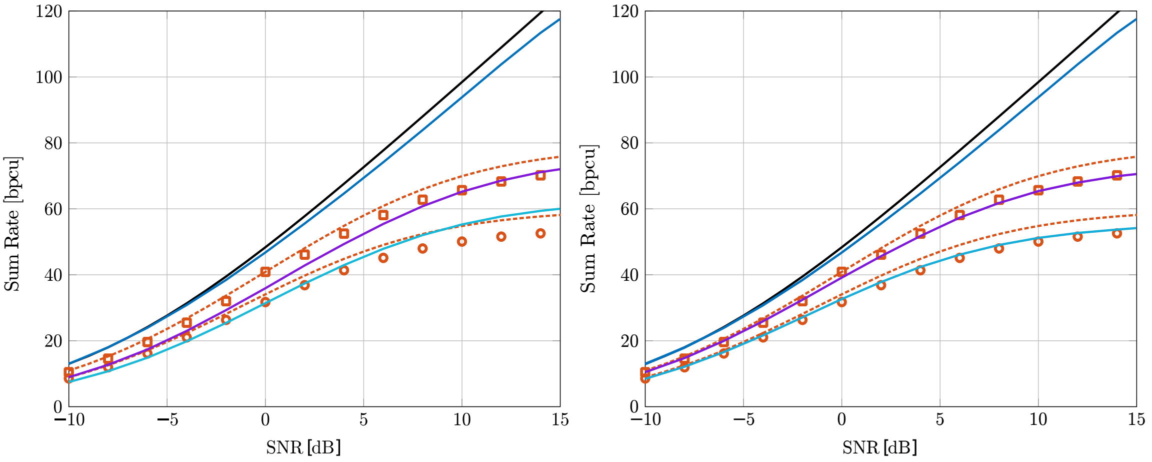
```
Inputs: $\boldsymbol{u}$, $\boldsymbol{H}$, $\mathcal{S} = \{1, \ldots, N\}$
Initialize: $\boldsymbol{x} = \boldsymbol{x}_{init}$, $\alpha = \alpha_{init}$
for $i = 1 : I$ do
    while error decreasing do
        for $n \in \mathbb{S}$ do
            $(x^{\star}_{n^{\star}}, n^{\star}) = \arg\min_{x_n \in \mathcal{X}, n \in \mathbb{S}} MSE(\boldsymbol{x}, \alpha)$
            $(x_1^{(i)}, \ldots, x_n^{(i)}, \ldots, x_{N-M}^{(i)})^{\mathrm{T}} = (x_1^{(i)}, \ldots, x^{\star}_{n^{\star}}, \ldots, x_{N-M}^{(i)})^{\mathrm{T}}$
            $\mathbb{S} \leftarrow \mathbb{S} \setminus \{n^{\star}\}$
        end
    end
    $\boldsymbol{x}_{FRF}^{(i)} = \boldsymbol{H}_{FRF}^{\mathrm{H}} \left(\boldsymbol{H}_{FRF}\boldsymbol{H}_{FRF}^{\mathrm{H}} + \gamma\boldsymbol{I}\right)^{-1} (\boldsymbol{u} - \alpha^{(i)}\boldsymbol{H}_{QRF}\boldsymbol{x}_{QRF}^{(i)})$
    $\boldsymbol{x}^{(i)} = [\boldsymbol{x}_{QRF}^{(i)},\ \boldsymbol{x}_{FRF}^{(i)}]$
    $\alpha^{(i)} = \frac{\mathrm{Real}\{\mathbf{u}^{\mathrm{H}}\mathbf{H}\mathbf{x}^{(i)}\}}{||\mathbf{H}\mathbf{x}^{(i)}||_2^2 + K\sigma^2}$
end
Output $\boldsymbol{x}, \alpha$
```

and uniform 256-QAM inputs is plotted to benchmark the quantized precoders.

We first note that the lower bounds obtained from the Bussgang decomposition in Fig. 3.18b employ closed form approximations for the covariance matrix $\boldsymbol{C}_d$ and Gaussian input distributions, and are thus evaluated with (3.31). The lower bounds based on histogram approximations of probability density functions, discussed in [65], are shown as a benchmark, and numerically evaluated for uniform 256-QAM inputs.

Second, as shown in Fig. 3.18a, the choice of quantizer (signaling alphabet) has a non-negligible impact on the achievable rates. The choice of $\mathcal{X}$ results in a loss of average transmit power due to the 0 symbol, and consequently a rate loss in the low SNR regime. However, in the high SNR regime, especially for the case with $b = 2$, it outperforms both the Bussgang approximation and the optimized quantization from [65], designed for Gaussian signals. This suggests that an optimization of the signaling alphabet is worthwhile.

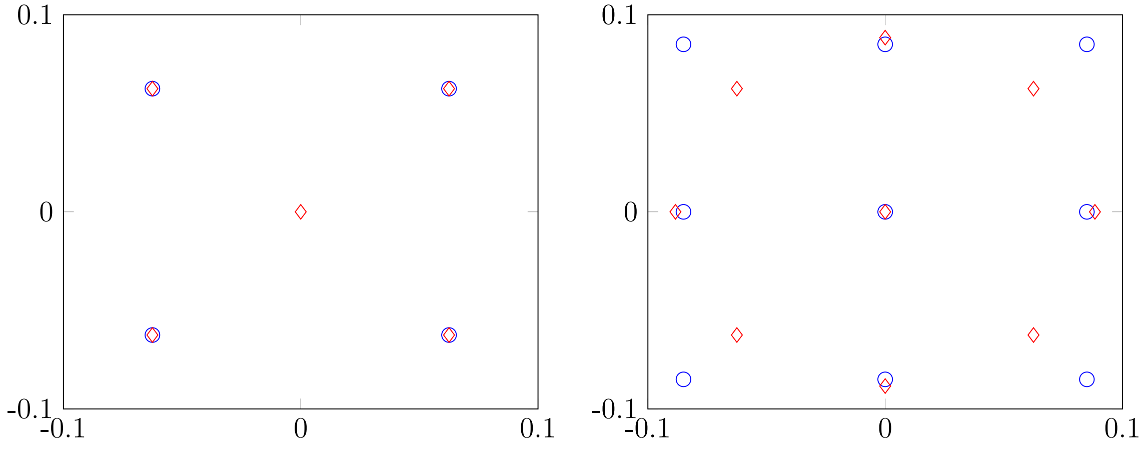
QLP is a competitive choice for low SNR, exhibiting the same computational complexity as classical LP schemes and achievable rates. However, at high SNR these information rates saturate and result in large penalties. This is the main motiva-



(a) GMI evaluated with the uniform quantizer $\mathcal{X}$

(b) GMI evaluated with the quantizer from [65]

Figure 3.18.: Achievable sum rates for Gaussian WF(—), 256-QAM WF(—), QLP with Bussgang approximation(- -) from [65], GMI with QLP and $b = 2$(—), GMI with QLP and $b = 3$(—), histogram-based numerical lower bounds from [65] for $b = 2$(○) and for $b = 3$(□)



(a) Signaling constellation $\mathcal{X}$ (◇) and the constellation from [65] (○) for $b = 2$.

(b) Signaling constellation $\mathcal{X}$ (◇) and the constellation from [65] (○) for $b = 3$.

Figure 3.19.: Signaling alphabets.

tion to develop computationally efficient non-linear methods that mitigate this loss in performance.

3.10.2. Non-linear Precoding

For the more advanced methods, we first show the results for flat fading channels. As a reference, we use the zero forcing (ZF) solution with infinite precision ADCs. We also include the performance of the MF and ZF QLP schemes, and the performance of the SQUID and ADMM algorithms. MAGIQ performs $I = 3$ iterations, SQUID $I = 50$ iterations, and ADMM uses $I = 100$ iterations. Going beyond these values did not result in further noticeable improvements in our scenarios. We do not claim that these numbers are optimal, but we tuned them for good performance for higher order modulations.

We apply the same α optimization as Scenario 1 to SQUID with $J = 4$ outer iterations. We call this precoding algorithm SQUID-α. For MAGIQ and ADMM, the quantization resolution is one bit per real dimension, i.e., we set $b = 2$ in (3.7). The precoding solution for SQUID uses phase modulation only, i.e., the zero symbol is not included in the transmitter alphabet. Also, MAGIQ is initialized with the solution of the quantized MF in order to speed up convergence. The results are obtained from $2,000$ channel realizations for flat fading, and 200 channel realizations for frequency selective fading.

We show the achievable rates for 16-QAM and 64-QAM in Figs. 3.20 and 3.21, respectively. Both MAGIQ and ADMM show similar performance for higher order modulation formats. For 16-QAM, SQUID-α with adaptive α is competitive with MAGIQ and ADMM, but exhibits convergence issues at high signal-to-noise ratio (SNR) for large spectral efficiency. QLP is shown here again in order to quantify the improvement offered by non-linear precoders. The gap becomes more pronounced as higher spectral efficiency is required.

Fig. 3.22 shows the achievable rates with 64-QAM in a block fading scenario with the transceivers designed according to the guidelines presented in Sec. 3.2.3. Observe that both Scenarios 2 and 3 seem to be interference limited at high SNR, but Scenario 3 operates close to Scenario 1. This demonstrates that broadcasting the precoding factors at the receiver is not necessary.

Fig. 3.23 shows the trade-off between achievable rates, number of antennas and resolution of the digitally controlled phase shifters. We first note that a system with

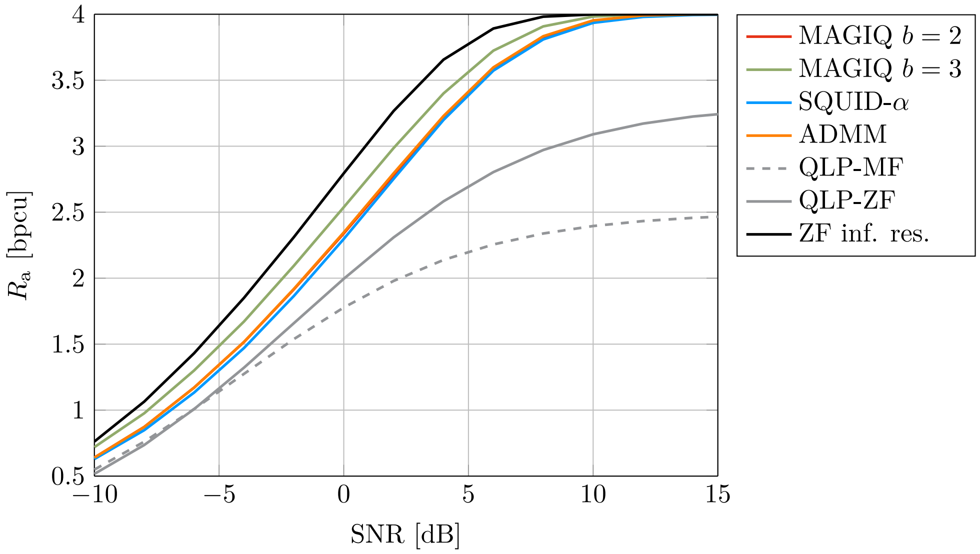


Figure 3.20.: GMI rates for 16-QAM, $N = 128$, and $K = 16$.

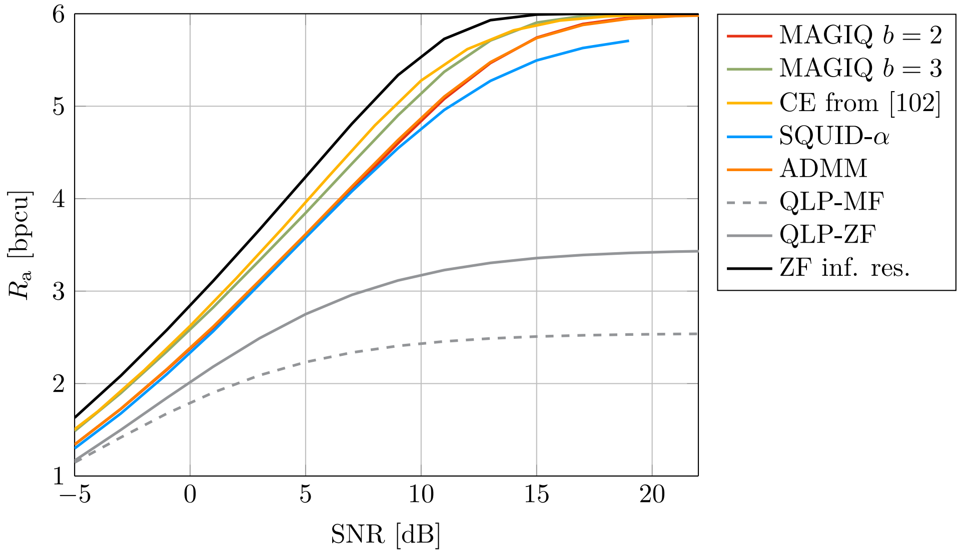


Figure 3.21.: GMI rates for 64-QAM, $N = 128$, and $K = 16$.

$b = 2$ bits of phase needs about 1.3 times more antennas than one with an infinite resolution ZF precoder to achieve the same average user rates. Second, we notice that increasing the number of bits from $b = 2$ to $b = 6$ has diminishing returns, and saturates away from the ZF even for an infinite number of bits, as noted before in our

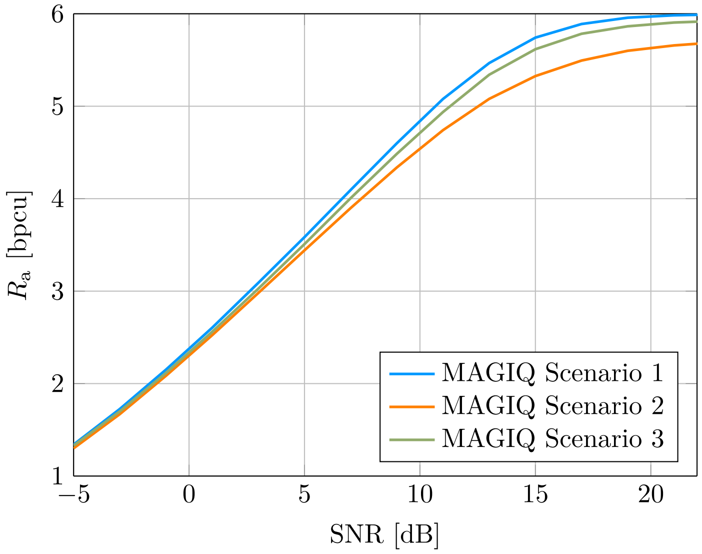


Figure 3.22.: GMI rates for 64-QAM, $N = 128$, and $K = 16$.

analysis of SDP-based upper and lower bounds. The setting with $b = 3$ yields 3/4 of the rate increase and all other settings with $b = 4, \ldots, \infty$ the remaining 1/4. This suggests that $b = 3$ offers a good trade-off between achievable rates, computational and hardware complexity.

Frequency selective fading is considered in Fig. 3.24 for a channel with $L = 15$ taps, $N = 128$ antennas, $K = 16$ users, and OFDM symbols with $T = 286$. We evaluate the GMI for 16-QAM and 64-QAM. The GMI reported is computed per-subcarrier and then averaged over the subcarriers. This assumes channel coding is performed over multiple subcarriers and symbols. The frequency selective MAGIQ performs $I = 4$ iterations for each OFDM symbol and computes a single precoding factor over the OFDM symbol. The algorithm was initialized with a time domain quantized solution of the frequency domain MF. The results show that the gap to the flat fading Scenario 1 solution is negligible. We further notice that QCM with 6 iterations is equivalent with MAGIQ, with only an insignificant SNR penalty.

Fig. 3.25 compares achievable rates of QCM, SQUID and MSM for a medium-sized scenario ($N = 64, K = 8$) borrowed from [75]. We choose M-PSK because the MSM algorithm was originally proposed for OFDM with PSK. We see that MSM is superior to both SQUID and QCM at lower SNR (and low-medium rates), whereas at high

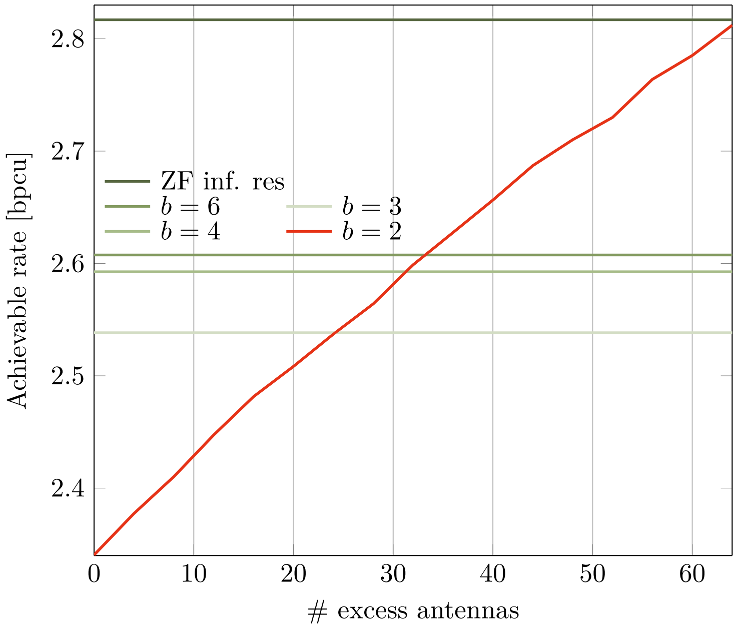


Figure 3.23.: Excess antennas vs. resolution for $K = 16$, $\mathrm{SNR} = 0\,\mathrm{dB}$, 16-QAM. The baseline systems (green) has $N = 128$ transmit antennas.

rates QCM is undoubtedly superior. We also tried to replicate our larger system scenario but unfortunately the MSM algorithm could not be run on our simulation platform because of memory limitations (2 AMD EPYC 7282 16-Core, 125GiB of system memory, Matlab with both dual-simplex and interior-point solvers).

Fig. 3.26a shows the impact that iterations have on the achievable rates for OFDM, as well as the convergence of the proposed MAGIQ and QCM. We observe that both the greedy and the cyclic/random schedules achieve very good performance although at the cost of extra iterations (5-6 compared to 3-4 for MAGIQ).

Fig. 3.26b shows that the inclusion of the precoder factor α in the optimization is fundamental to the good performance of MAGIQ, exhibiting a gap as large as 10dB to the precoder without the inclusion of a precoding factor.

Finally, we show simulations for a Winner 2 non-line-of-sight (NLOS) C2 urban scenario [105]. The purpose of this investigation is to confirm the viability of the proposed precoders in semi-deterministic channel models that are extensively used as proxies for real physical channels. The considered geometric layout is shown in Fig. 3.27 and the detailed parameters are as follows:

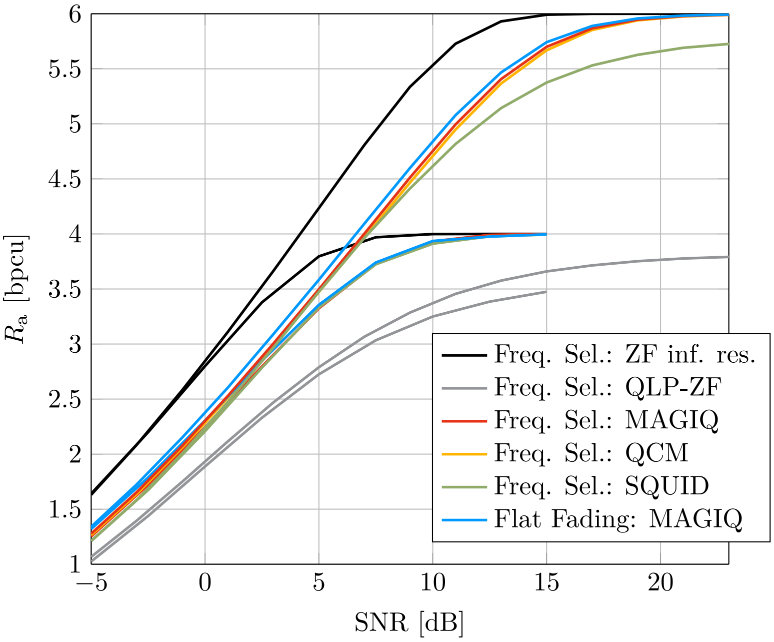


Figure 3.24.: GMI rates for OFDM, 16-QAM and 64-QAM, $N = 128$, $K = 16$, $T = 286$, $L = 15$.

- ▷ 8 uniformly distributed users on a disk of radius 150m at every drop;
- ▷ 80 (8x10) dipole uniform rectangular antenna array at the base station (with $\lambda/2$ spacing);
- ▷ 5 MHz bandwidth at a 2.53 GHz center frequency;
- ▷ MAGIQ precoder with $b = 2$ bits of phase;
- ▷ OFDM with 128 subcarriers and uniform 16-QAM;
- ▷ No Doppler shift, shadowing and pathloss;
- ▷ we compare with a Rayleigh 22 tap channel.

Fig. 3.28 shows the achievable GMI rates for both an infinite precision ZF precoder and the MAGIQ frequency selective precoder. For a Winner 2 NLOS scenario, two-bit MAGIQ with 16-QAM performs close to its Rayleigh counterpart. At high SNR, there is a decrease in the GMI slope as compared to ZF. This shows that for more correlated channels one may need to increase the number of bits (or equivalently the number of antennas) to avoid becoming interference limited.

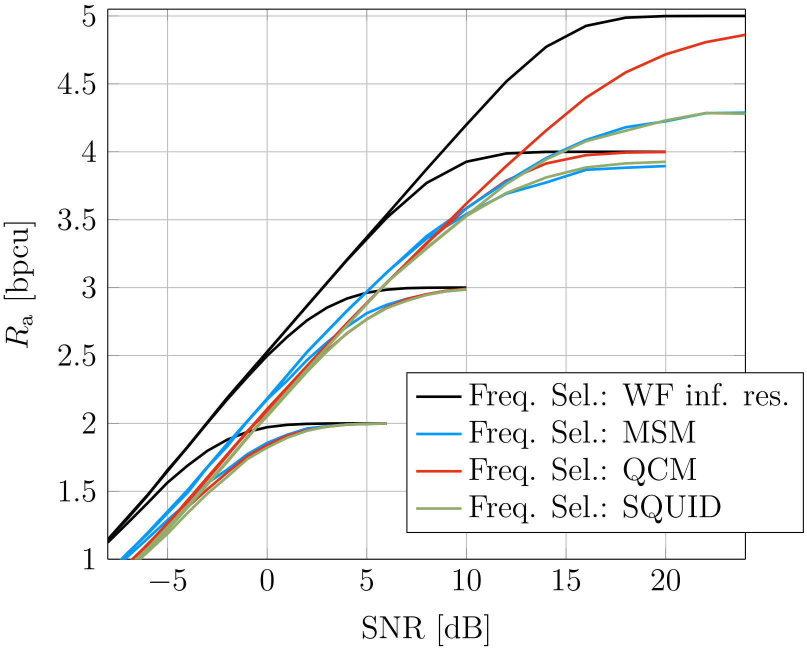
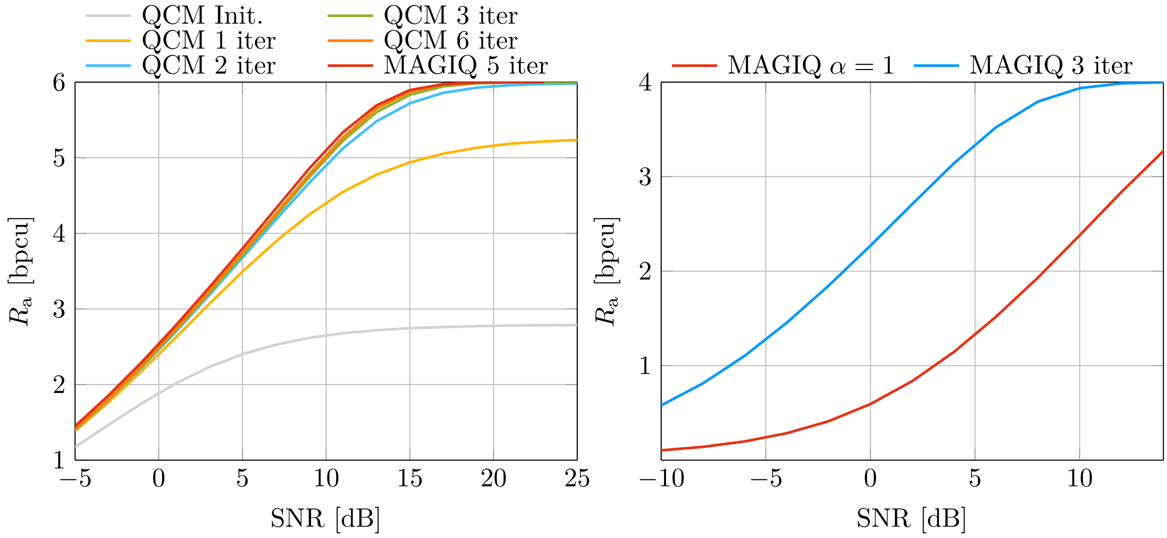


Figure 3.25.: GMI rates for OFDM with 4, 8, 16, 32-PSK, 2 bits, $N = 64$, $K = 8$, $T = 32$, $L = 4$.

3.10.3. Mixed Resolution RF chains

Recall that FRF refers to high resolution RF chains and QRF to low resolution RF chains. Fig. 3.29 shows the GMI for a Rayleigh flat fading channel where the transmitter has $N = 32$ antennas of which $M = 16$ or $M = 12$ are QRF antennas. The QRF antennas use two bits of phase $b = 2$. We further consider 64-QAM modulation and $K = 16$ UEs. The performance is compared with a WF with infinite resolution with $N = 32$, and also with plain MAGIQ ($M = 0$) with $N = 32$. Observe that MAGIQ cannot support high order modulations (not even 16-QAM) for such a system load, whereas the FRF-MAGIQ can achieve the required spectral efficiency with a SNR penalty compared to WF. Alt-MAGIQ has convergence issues at high SNR, i.e., it can get trapped in poor local minima.

Doubling the number of antennas to $N = 64$ considerably improves the GMI of plain MAGIQ with $b = 2, 3$, as shown in Fig. 3.30. Observe that both FRF-MAGIQ and Alt-MAGIQ with $N - M = K$ have convergence issues in the high SNR regime. The



(a) GMI rates vs Iterations for OFDM, 64-QAM, for QCM with $N = 128$, $K = 16$, $T = 286$, $L = 15$ and $b = 3$.

(b) GMI rates with/without α for Flat Fading for 16-QAM, $N = 128$, and $K = 16$.

Figure 3.26.: Impact of iterations and optimization of α for MAGIQ for Flat and Frequency Selective Channels.

number of iterations for FRF-MAGIQ and Alt-MAGIQ is fixed at four. However, the computational complexity of the two algorithms is vastly different because of the different update schedules. Alt-MAGIQ esentially requires approximately four times the number of computations of MAGIQ and one WF computation (the WF is computed only once for all iterations and is reused), whereas FRF-MAGIQ computes $(N - M)|\mathcal{X}|$ WF solutions (a matrix vector multiplication) at each iteration.

3.11. Conclusions

We introduced QCM and MAGIQ for quantized precoding in a massive MIMO downlink channel. MAGIQ applies to both frequency-flat and frequency selective channels. Numerical performance comparisons using lower bounds on information rates suggest that MAGIQ outperforms QLP, SQUID and MSM, and it achieves similar performance as ADMM. We studied different update schedules of the precoding factor α. For frequency-selective channels and OFDM, MAGIQ loses only about 0.25 dB as

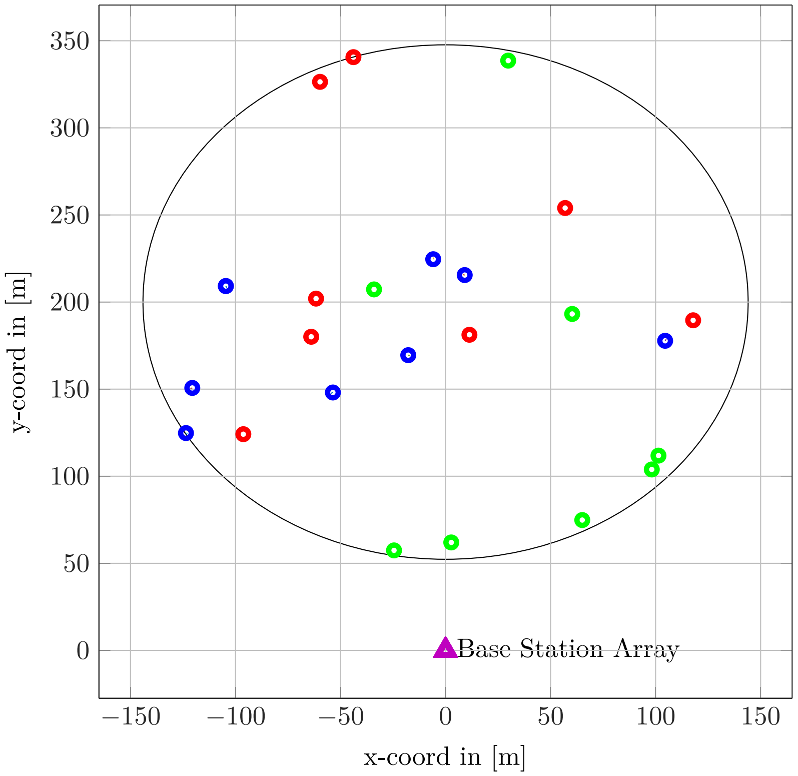


Figure 3.27.: Single-cell Winner 2 network scenario with 3 independent drops of 8 users shown in red, green and blue.

compared to the frequency-flat case for 16-QAM at 3 bpcu. The results further show that the implementation complexity of MAGIQ can be significantly reduced through slight modifications in the optimization strategy for frequency selective channels without sacrificing achievable rates.

MAGIQ and QCM with very low numbers of bits $b = 2$ or $b = 3$ with a relatively high system load ($K/N \approx 0.5$) are insufficient to achieve high spectral efficiency. However, a combination of high resolution and low resolution RF chains with precoding algorithms such as Alt-MAGIQ or FRF-MAGIQ recover the performance observed with much larger systems with lower system loads, at affordable computational complexity.

Future work should consider the combination of MAGIQ with multi-user scheduling and the influence of imperfect CSI at the receiver. Another research direction is

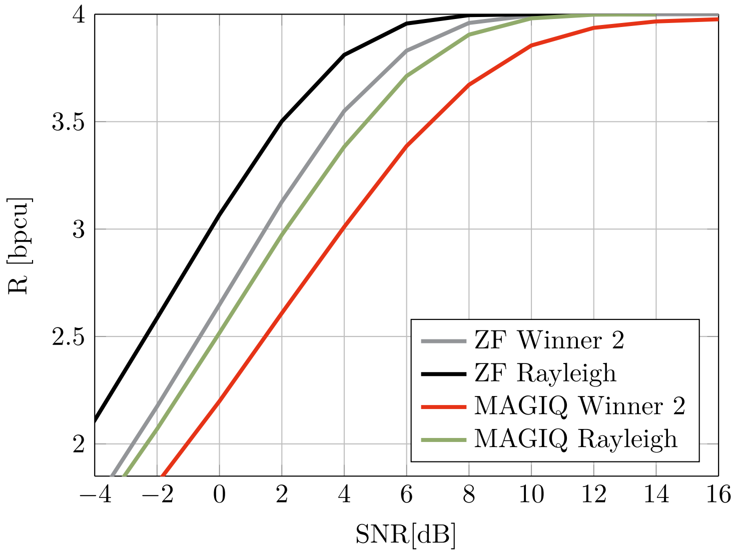


Figure 3.28.: GMI rates for OFDM, 16-QAM, $N = 80$, $K = 8$, $T = 286$, Winner 2 and Rayleigh channels

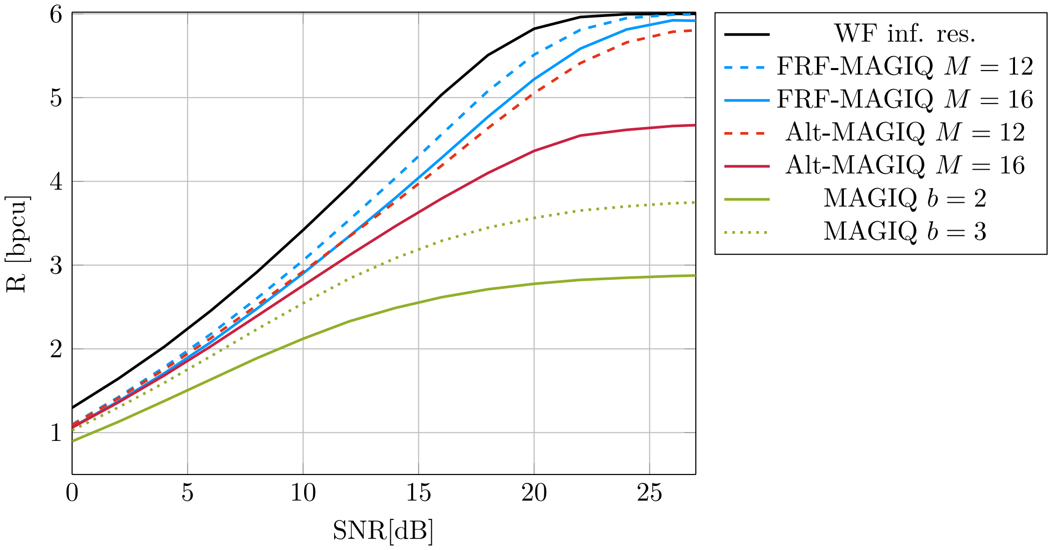


Figure 3.29.: GMI rates for 64-QAM, $N = 32$, $M = 16\,\&\,12$, $K = 16$, $b = 2$

to extend MAGIQ and QCM for multiple antennas at the receivers. It is also of practical importance to analyze a transmission chain where both the transmitter and the receivers operate at very low resolutions. Finally, one should evaluate the

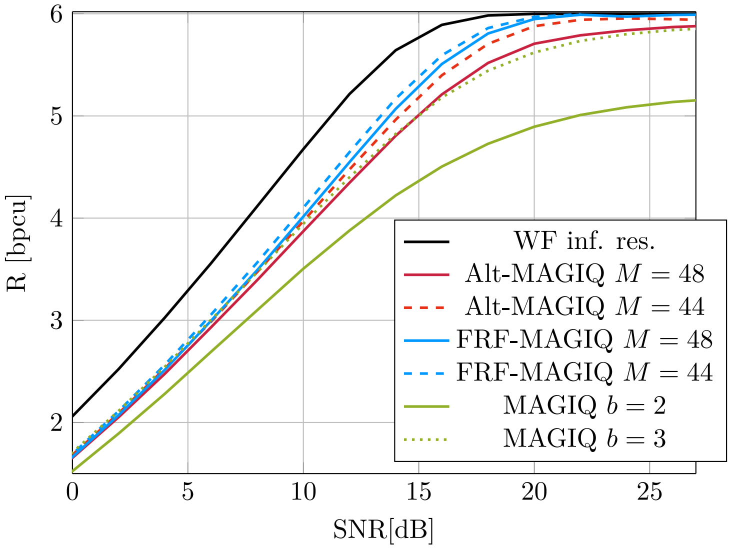


Figure 3.30.: GMI rates for 64-QAM, $N = 64$, $M = 48\,\&\,44$, $K = 16$, $b = 2$

performance of the proposed algorithms in combination with time and frequency synchronization at the receiver.

4

Summary and Outlook

MIMO communications are essential to achieve the high spectral efficiency targets of future communications standards. However, integrating more antennas in the limited space of handheld devices, or scaling up the number of base station antennas into the hundreds, can be complex and expensive. This thesis investigated two aspects of MIMO systems and proposed information-theoretic and hardware-informed approaches for both compact and massive MIMO.

4.1. Summary

Chapter 2 presented a circuit-level model of MIMO RF front-ends that captures the physical interactions of signal and noise caused by antenna array elements operating in each others' electromagnetic fields. We reviewed a closed-form optimal design for lossless passive matching networks that maximizes the mutual information rates of a Gaussian linear channel. The solution is sensitive to manufacturing variations that are present in real-world implementations of such a circuit. The effects are shown to be a function of the matching network architecture and the antenna array design, suggesting that substantial performance degradation could be mitigated by appropriate design choices. The optimal matching networks are quintessentially narrow-band and are no longer optimal outside a small bandwidth centered at the carrier frequency.

In Chapter 3 we studied downlink precoding for multiuser MIMO with large antenna arrays at the base station and with frequency selective fading. To make such systems practical, scalable and power efficient, the transmitted signals were restricted to have a constant envelope and, in addition, to have a low resolution compared to existing deployments. The constant envelope limitation was motivated by efficient power amplifier design, while coarse quantization relaxes the design of DACs and considerably reduces their power consumption. Simple implementations of low resolution quantization of classical linear precoders result in significant performance degradation at high SNR. We showed that a computationally efficient nonlinear precoder can recover most of the loss and operate close to the information rates of infinite resolution MMSE precoders. Finally, we showed that the proposed architecture and algorithms can be seamlessly integrated with existing infrastructure for energy and cost efficient upgrades to Massive MIMO.

4.2. Outlook

Future Massive MIMO base stations will need to be compact for dense deployment. Thus merging the two topics studied in this thesis is a natural progression of research. An example is offered in [106] where sigma-delta quantizers are considered together with closely spaced antennas to direct distortion caused by coarse quantization away from intended directions of propagation (position of intended users).

An interesting direction to expand the results in Chapter 3 is to develop realistic channel estimation error models when the same array is used to estimate the uplink CSI. These models could lead to robust precoding schemes in combination with receiver design. For example, supervised learning techniques such as autoencoders could learn the models implicitly from data by adapting parameters of neural networks to minimize suitably chosen cost functions.

Next, in Chapter 2 we showed that the information rates for coupled MIMO arrays depend on the matching networks and the shape of the antenna array. We optimized the matching networks, but did not carefully consider the antenna array. An interesting direction of research is to optimize the positions and/or the types of antenna

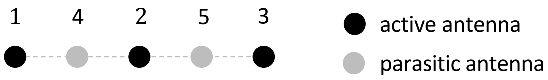


Figure 4.1.: ULA with 3 active and 2 parasitic elements

elements. Figure 4.1 shows an example: instead of changing the positions or the design of the antenna elements we could choose a uniform spacing but place antenna elements that are simply terminated by a single impedance between antenna elements connected to the (low noise) amplifiers. These elements are called parasitic elements in the antenna and propagation literature [107].

Figure 4.2 shows information rates vs. antenna separation between antenna elements connected to low noise amplifiers. In solid red we show the rates of a 5x3 MIMO link without parasitic elements and in solid green we show the rates for the same 5x3 MIMO for self matching. In solid blue we show the rates achieved with parasitic elements at the receiver, i.e., we are using 5x(3+2) MIMO where the matching network connects 3 antenna elements to low noise amplifiers. The difference from self matching is that we use a genetic algorithm to jointly optimize the diagonal matching elements and the impedances that terminate the parasitic antenna elements. As shown in the Figure, this procedure achieves information rates as good as the optimal system without parasitic elements. We conclude that a joint design of the antenna array and the matching network is an interesting approach to simplify matching networks without sacrificing performance.

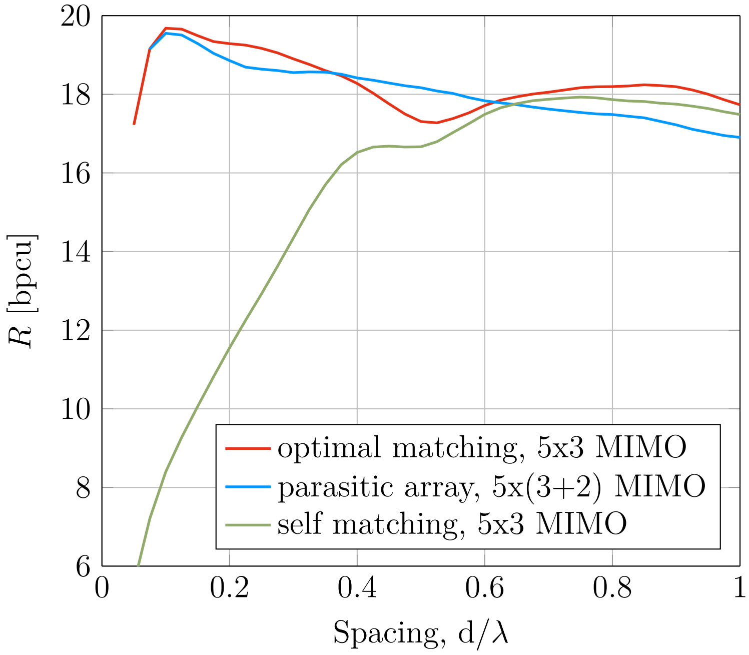


Figure 4.2.: ULA with parasitic antenna elements

A

Mismatched Decoding and Information Rates

This appendix reviews information theoretic arguments that show that the GMI is a lower bound to the information rate. The GMI thus has an operational meaning as an *achievable rate*: there exist codes approaching this rate and decoders for which the error probability approaches zero as the block length increases.

Capacity

Shannon's coding theorem [108] provides conditions under which reliable communication is possible over noisy channels. Central to the achievability proof is knowing the channel law at both ends of the communication chain. However, in practice the channel law may not be known, or may be available only partially. We focus on the point-to-point coding problem with a mismatched decoder.

The input and output alphabets are finite and denoted as $\mathcal{X}$ and $\mathcal{Y}$, respectively. The probability density of the channel output sequence $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]$ given the input sequence $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$ is $p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} p_{Y|X}(y_i|x_i)$.

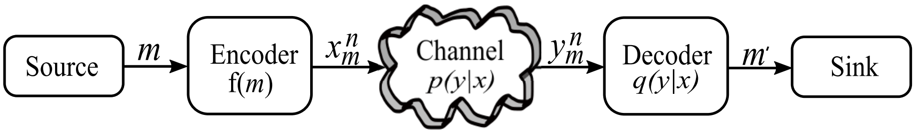


Figure A.1.: Information theoretic model of a digital communication system with mismatched decoders

The encoder is a mapping $f(\cdot)$ from a set $\{1, \dots, M\}$ of equal probability messages. The codeword $\boldsymbol{x}_m$ for message m is chosen from the codebook $\mathcal{C} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_M\}$. The receiver estimates the transmitted message m' from the received sequence $\boldsymbol{y}$:

$$m' = \arg \max_{j \in \{1, \dots, M\}} q(\boldsymbol{y}|\boldsymbol{x}_j), \quad \text{with } q(\boldsymbol{y}|\boldsymbol{x}_j) = \prod_{i=1}^{n} q(y_i|x_{j,i}) \tag{A.1}$$

where $q(y|x)$ is a non-negative mismatched decoding metric, i.e. a function $q : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$. We note that the metric $q(y|x)$ need not be a probability density function, i.e. we do not require $\sum_y q(y|x) = 1$. However, we require $q(y|x) \geq 0$ for all x, y.

We define an average error probability associated with the codebook $\mathcal{C}$: $\bar{P}_e(\mathcal{C}) = \Pr(m \neq m')$. A rate $R = \frac{1}{n} \log M$ is said to be achievable if, for all $\epsilon > 0$, there is a codebook $\mathcal{C}_n$ of length n with $M \geq e^{n(R-\epsilon)}$ codewords under the decoding criteria given in A.1.

When $q(\boldsymbol{y}|\boldsymbol{x}) = p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})$, we obtain the maximum likelihood decoder which is analyzed in Chapter 5 of [109] and shown to achieve the Shannon capacity:

$$\begin{aligned} C &= \max_{P_X} I(X;Y) \\ I(X;Y) &= \sum_{x,y} P_X(x)\, p_{Y|X}(y|x) \log \frac{p_{Y|X}(y|x)}{p_Y(y)} \end{aligned} \tag{A.2}$$

where $p_Y(y) = \sum_{a \in \mathcal{X}} P_X(a)\, p_{Y|X}(y|a)$ for all $y \in \mathcal{Y}$.

Mismatched Decoding Rates

We follow the techniques from [109, Ch. 5] to derive an error exponent for the mismatched DMC,that recovers the GMI [62].

Theorem A.1. (see [109, Problem 5.22]) For a DMC $p_{Y|X}(y|x)$, the decoder metric $q(y|x)$, an input distribution $P_X(x)$ and $s \geq 0$, the following rate is achievable with the i.i.d. random coding ensemble:

$$\begin{aligned} C_{\mathrm{GMI}}(p_{Y|X}, q, P_X) &= \max_{P_X} \mathrm{I}_{\mathrm{GMI}}(X;Y) \\ I_{\mathrm{GMI}}(X;Y) &= \sup_{s\geq 0} \sum_{x,y} P_X(x) p_{Y|X}(y|x) \log \frac{q(y|x)^s}{q(y)} \end{aligned}$$

where $q(y) = \sum_{a\in\mathcal{X}} P_X(a) q(y|a)^s$ for all $y \in \mathcal{Y}$.

We first show that I_{GMI} is a lower bound to A.2 [110]:

$$\begin{aligned} &\mathrm{I}(X;Y) - \mathrm{E}\left[\log_2\left(\frac{q_{Y|X}(Y|X)^s}{q(Y)}\right)\right] \\ &= \sum_x \sum_y P_X(x) p_{Y|X}(y|x) \left[\log \frac{P_X(x) p_{Y|X}(y|x)}{P_X(x) p_Y(y)} - \log \frac{q_{Y|X}(y|x)^s}{q(y)}\right] \mathrm{d}y \\ &\overset{(a)}{=} \boldsymbol{D}\Big(P_X p_{Y|X} \Big\| \underbrace{\frac{P_X q^s_{Y|X}}{q}}_{t_{X|Y}} p_Y\Big) \\ &\geq 0 \end{aligned} \tag{A.3}$$

where (a) follows because $t_{X|Y}(\cdot|y)$ is a distribution: $\sum_{x\in\mathcal{X}} t_{X|Y}(x|y) = 1$ and where the inequality follows from the non-negativity of the KL divergence, and becomes equality if $q^s_{Y|X} = c\, p_{Y|X}$, where c is a constant. An alternative derivation with error exponents can be found in [62].

The i.i.d. ensemble is defined in the literature [109] as the set of codebooks where each symbol in each codeword is generated independently according to $Q_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^n P_X(x_i)$. We start from the average error probability:

$$\mathrm{Pr}(m \neq m') = \sum_{m\in\{1,\ldots,M\}} \mathrm{Pr}(m \neq m'|m)\mathrm{Pr}(m) \tag{A.4}$$

$$= \sum_{\boldsymbol{x}_m} \sum_{\boldsymbol{y}} Q_{\boldsymbol{X}}(\boldsymbol{x}_m) p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}_m) \mathrm{Pr}(m' \neq m|m, \boldsymbol{x}_m, \boldsymbol{y}) \tag{A.5}$$

where we are abusing notation by considering m and m' as both random variables and realizations of random variables, the meaning being clear from the context. The probability or error is summed over all possible selections of the message $\boldsymbol{x}_m$ corresponding to a message m, and all the sequences $\boldsymbol{y}$ at the output of the channel.

We take a closer look at how to determine a decoding error event. Given $m, \boldsymbol{x}_m$ and $\boldsymbol{y}$ then the decoder will decide for the wrong message $\tilde{m}$ if $q(\boldsymbol{y}|\boldsymbol{x}_{\tilde{m}}) \geq q(\boldsymbol{y}|\boldsymbol{x}_m)$. Thus:

$$\begin{aligned}
\Pr(m' \neq m|m, \boldsymbol{x}_m, \boldsymbol{y}) &= \Pr\left(\bigcup_{m \neq m'} \frac{q(\boldsymbol{y}|\boldsymbol{X}_{\tilde{m}})}{q(\boldsymbol{y}|\boldsymbol{x}_m)} \geq 1\right) && \text{(A.6)}\\
&\leq \sum_{m \neq m'} \Pr\left(\frac{q(\boldsymbol{y}|\boldsymbol{X}_{\tilde{m}})}{q(\boldsymbol{y}|\boldsymbol{x}_m)} \geq 1\right) && \text{(A.7)}\\
&\leq \sum_{m \neq m'} \left(\Pr\left(\frac{q(\boldsymbol{y}|\boldsymbol{X}_{\tilde{m}})^s}{q(\boldsymbol{y}|\boldsymbol{x}_m)^s} \geq 1\right)\right)^\rho, \quad 0 < \rho \leq 1, \quad s \geq 0 && \text{(A.8)}\\
&\leq \left(\sum_{m \neq m'} \frac{\mathrm{E}\left[q(\boldsymbol{y}|\boldsymbol{X}_{\tilde{m}})^s\right]}{q(\boldsymbol{y}, \boldsymbol{x}_m)^s}\right)^\rho, \quad 0 < \rho \leq 1, \quad s \geq 0 && \text{(A.9)}
\end{aligned}$$

where the last inequality follows by the Markov inequality.

Substituting back and considering that $m \neq m'$ can happen in $M-1$ ways:

$$\Pr(m' \neq m|m, \boldsymbol{x}_m, \boldsymbol{y}) \leq \left((M-1)\frac{\mathrm{E}\left[q(\boldsymbol{y}|\boldsymbol{X}_{\tilde{m}})^s\right]}{q(\boldsymbol{y}|\boldsymbol{x}_m)^s}\right)^\rho. \tag{A.10}$$

For a DMC we have $q(\boldsymbol{y}|\boldsymbol{x}_m) = \prod_{i=1}^n q(y_i|x_{m,i})$ and $p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}_m) = \prod_{i=1}^n p_{Y|X}(y_i|x_{m,i})$ and the right hand side of equation A.10 is:

$$\left((M-1)\frac{\mathrm{E}\left[q(\boldsymbol{y}|\boldsymbol{X}_{\tilde{m}})^s\right]}{q(\boldsymbol{y}|\boldsymbol{x}_m)^s}\right)^\rho = (M-1)^\rho \prod_{i=1}^n \left(\frac{\mathrm{E}\left[q(y_i|X_{\tilde{m},i})^s\right]}{q(y_i|x_{m,i})^s}\right)^\rho. \tag{A.11}$$

Finally, the average error probability is:

$$\begin{aligned}
\Pr(m \neq m') &\leq (M-1)^\rho \prod_{i=1}^n \sum_x \sum_y P_X(x_i) p_{Y|X}(y_i|x_i) \left(\frac{q(y_i)}{q(y_i|x_i)^s}\right)^\rho && \text{(A.12)}\\
&\leq 2^{nR\rho} \left(\sum_x \sum_y P_X(x) p_{Y|X}(y|x) \left(\frac{q(y)}{q(y|x)^s}\right)^\rho\right)^n && \text{(A.13)}
\end{aligned}$$

$$= 2^{-n(f(q,p_{Y|X},\rho,s)-R\rho)} \tag{A.14}$$

where

$$f(q,p_{Y|X},\rho,s) = -\log_2\left(\sum_x\sum_y P_X(x)p_{Y|X}(y|x)\left(\frac{q(y)}{q(y|x)^s}\right)^\rho\right). \tag{A.15}$$

Any rate that leads to a non negative $f(q,p_{Y|X},\rho,s) - R\rho$ is achievable, i.e. the average error probability for these rates vanishes exponentially fast as $n \to \infty$. We find the tightest bound by optimizing over ρ and s. According to [109] a stationary point of $f(q,p_{Y|X},\rho,s) - R\rho$ is also a maximizer of the exponent. We thus have

$$R = \left.\frac{\partial f(q,p_{Y|X},\rho,s)}{\partial\rho}\right|_{\rho=0} = \mathrm{E}\left[\log_2\left(\frac{q(Y|X)^s}{q(Y)}\right)\right] \tag{A.16}$$

and obtain the GMI capacity as

$$C_{\text{GMI}} = \max_{P_X}\sup_{s\geq 0}\mathrm{E}\left[\log_2\left(\frac{q(Y|X)^s}{q(Y)}\right)\right]. \tag{A.17}$$

This concludes the achievability proof for the GMI. Although presented for finite alphabet DMCs, the results can be extended to infinite output alphabets and share almost identical proofs.

Finally, note that $q(y|x)$ is a symbol-wise metric operating on symbols from $\mathcal{X}$. When binary codes are combined with high order modulation a common choice is a bit-wise metric $q(y|x) = \prod_{j=1}^m q(y|b_j)$, where the labeling function $\Pi : \mathcal{X} \to \{0,1\}^m$ maps symbols from $\mathcal{X}$ onto their m-dimensional binary labels $\Pi(x) = b_1,\ldots,b_m$. Suppose in addition that the random experiment samples codewords from $P_X(x) = \prod_{j=1}^m P_{B_j}(b_j)$. Note that the individual bit levels B_j can be individually probabilistically shaped. Similar to (A.4-A.18) we obtain:

$$C_{\text{GMI}} = C_{\text{BICM}} = \sum_{j=1}^m I(B_j;Y) \tag{A.18}$$

which is the bit interleaved coded modulation capacity (BICM) [63]. The operational meaning of the GMI, the existence of a (random) code that is capacity achieving, and its connection to BICM motivated our approach to cast the quantized precoding problem into the GMI framework.

B

SDP and Combinatorial Optimization

Semidefinite Programming (SDP) includes a large family of convex programming problems. Applications of SDPs include and are not limited to: control, combinatorial optimization, precoder design in communications, options pricing in finance, circuit synthesis and modeling in analog hardware design, robust and stochastic optimization, phase retrieval and matrix completion in machine learning, graph coloring and satisfiability problems in computer science. We are interested in the combinatorial problem presented in Sec. 3.2.1.

One paper that sparked a renewed interest in SDP for discrete problems is [111]. The authors propose an algorithm for the MAX-CUT problem (the task of finding an edge cut in a graph that contains the maximum number of edges). This is the first approximation algorithm based on a SDP relaxation of the original combinatorial problem. An approximation algorithm for the NP-hard MAX-CUT is a polynomial-time algorithm that computes a solution with some guaranteed quality for every instance of the problem. The guarantee in this case is in the form of an approximation ratio (for a graph G where $\mathbf{Algo}(G)$ is the value of the proposed solution and $\mathbf{OPT}(G)$ is the value of the optimal solution, the approx. ratio is $\alpha_{GW} = \inf_G \frac{\mathbf{Algo}(G)}{\mathbf{OPT}(G)} = 0.878$).

Note that if the approximation factor is equal to 1 then there is no gap and the relaxation can recover the solution to the initial problem.

We illustrate the concepts with a problem that is both NP-hard to solve exactly and also to approximate [112]:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \boldsymbol{x}^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{x} + 2\boldsymbol{q}^{\mathrm{T}} \boldsymbol{x} \\ \text{s.t.} \quad & \boldsymbol{x} \in \mathbb{Z}^n \end{aligned} \tag{B.1}$$

where $\boldsymbol{Q}$ is a symmetric positive semidefinite matrix and $\boldsymbol{q} \in \mathbb{R}^n$.

A number of relevant practical problems can be reduced to (B.1), including integer least squares, the closest vector and shortest vector problems. Note that every entry x_i in $\boldsymbol{x}$ is an integer and therefore satisfies $x_i \leq 0$ or $x_i \geq 1$, which can be equivalently written as quadratic constraint $x_i(x_i - 1) \geq 0$. As a result, (B.1) can relaxed as

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \boldsymbol{x}^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{x} + 2\boldsymbol{q}^{\mathrm{T}} \boldsymbol{x} \\ \text{s.t.} \quad & x_i(x_i - 1) \geq 0, \ i = 1, 2, \ldots, n. \end{aligned} \tag{B.2}$$

We now introduce a new variable $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}$ and a new constraint $\operatorname{diag}(\boldsymbol{X}) \geq \boldsymbol{x}$. These constraints can be seen as a reformulation of $x_i(x_i - 1) \geq 0$. We can thus re-write (B.1) as

$$\begin{aligned} \min_{\boldsymbol{X}, \boldsymbol{x}} \quad & \operatorname{Trace}(\boldsymbol{Q}\boldsymbol{X}) + 2\boldsymbol{q}^{\mathrm{T}} \boldsymbol{x} \\ \text{s.t.} \quad & \operatorname{diag}(\boldsymbol{X}) \geq \boldsymbol{x} \\ & \boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}. \end{aligned} \tag{B.3}$$

The problem is still not convex because of the constraint $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}$, which can be further relaxed to a convex form $\boldsymbol{X} \succeq \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}$. Finally we obtain the desired SDP relaxation of our initial problem:

$$\begin{aligned} \min_{\boldsymbol{X}, \boldsymbol{x}} \quad & \operatorname{Trace}(\boldsymbol{Q}\boldsymbol{X}) + 2\boldsymbol{q}^{\mathrm{T}} \boldsymbol{x} \\ \text{s.t.} \quad & \operatorname{diag}(\boldsymbol{X}) \geq \boldsymbol{x} \\ & \boldsymbol{X} \succeq \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}. \end{aligned} \tag{B.4}$$

We now introduce the randomization procedure for semidefinite relaxations. Suppose we instead solve (B.2) on average:

$$
\begin{aligned}
\min_{\tilde{\boldsymbol{X}}\sim\mathcal{N}(\nu,\Sigma)} \quad & \mathrm{E}\left[\tilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{Q}\tilde{\boldsymbol{X}} + 2\boldsymbol{q}^{\mathrm{T}}\tilde{\boldsymbol{X}}\right] \\
\text{s.t.} \quad & \mathrm{E}\left[\tilde{X}_i(\tilde{X}_i - 1)\right] \geq 0,\ i = 1, 2, \ldots, n.
\end{aligned}
\tag{B.5}
$$

which is equivalent to [113]:

$$
\begin{aligned}
\min_{\boldsymbol{\nu},\boldsymbol{\Sigma}} \quad & \mathrm{Trace}(\boldsymbol{\Sigma}\boldsymbol{Q} + \boldsymbol{\nu}\boldsymbol{\nu}^{\mathrm{T}}\boldsymbol{Q}) + 2\boldsymbol{q}^{\mathrm{T}}\boldsymbol{\nu} \\
\text{s.t.} \quad & \Sigma_{i,i} + \nu_i^2 - \nu_i \geq 0,\ i = 1, 2, \ldots, n
\end{aligned}
\tag{B.6}
$$

Problem (B.6) is in the form of (B.3) with $\boldsymbol{X} = \boldsymbol{\Sigma} + \boldsymbol{\nu}\boldsymbol{\nu}^{\mathrm{T}}$ and $\boldsymbol{x} = \boldsymbol{\nu}$.

Intuitively, the distribution $\mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Sigma})$ has a mean close to the least squares solution of (B.1) where the integer constraint is relaxed to the set of real numbers. In addition, the diagonal values of $\boldsymbol{\Sigma}$ are large enough such that $x_i(x_i - 1) \geq 0$ holds on average. Of course, samples from $\mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Sigma})$ are not always feasible for (B.2), but that can be resolved by projecting on the constraint space. The SDP thus provides a lower bound solution to the integer least squares problem, while the randomization procedure with projection gives an upper bound. In practice it was observed that the bounds are often tight and can be further tightened with additional constraints.

In this Appendix we followed the examples and exposition given in [113]. Although not explored here, SDP relaxation is an important foundation to building powerful algorithms such as branch-and-bound or cut-and-bound that can find good solutions to several classes of nonlinear integer and mixed-integer problems [114].

Bibliography

[1] ericsson.com, “Ericsson Mobility Report,” https://www.ericsson.com/4adc87/assets/local/mobility-report/documents/2020/november-2020-ericsson-mobility-report.pdf, 2020, [Online; accessed March 2021].

[2] M. J. Gans, “Channel capacity between antenna arrays—part II: amplifier noise dominates,” *IEEE Trans. Commun.*, vol. 54, no. 11, pp. 1983–1992, 2006.

[3] M. L. Morris and M. Jensen, “Network model for mimo systems with coupled antennas and noisy amplifiers,” *IEEE Trans. Antennas Propag.*, vol. 53, no. 1, pp. 545–552, 2005.

[4] C. P. Domizioli and B. L. Hughes, “Front-end design for compact MIMO receivers: a communication theory perspective,” *IEEE Trans. Commun.*, vol. 60, no. 10, pp. 2938–2949, 2012.

[5] D. Nie and B. M. Hochwald, “Broadband matching bounds for coupled loads,” *IEEE Trans. Circuits Syst. I*, vol. 62, no. 4, 2015.

[6] L. Kundu and B. L. Hughes, “Enhancing capacity in compact MIMO-OFDM systems with frequency-selective matching,” *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4694–4703, 2017.

[7] B. Debaillie, C. Desset, and F. Louagie, “A flexible and future-proof power model for cellular base stations,” in *IEEE Vehic. Technol. Conf. (VTC Spring)*, 2015, pp. 1–7.

[8] V. Venkateswaran and A.-J. van der Veen, “Analog beamforming in mimo communications with phase shift networks and online channel estimation,” *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4131–4143, 2010.

[9] F. Sohrabi and W. Yu, “Hybrid digital and analog beamforming design for large-scale antenna arrays,” *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 501–513, 2016.

[10] F. Steiner, “Coding for higher-order modulation and probabilistic shaping,” Dr.-Ing. Dissertation, Technische Universität München, 2020.

[11] H. Rothe and W. Dahlke, “Theory of noisy fourpoles,” *Proc. IRE*, vol. 44, pp. 811–818, Jun. 1956.

[12] M. Ivrlac and J. Nossek, “Toward a circuit theory of communication,” *IEEE Trans. Circuits Syst. I*, vol. 57, no. 7, pp. 1663–1683, July 2010.

[13] D. Pozar, *Microwave engineering.* Wiley, 2009.

[14] T.-K. Nguyen, C.-H. Kim, G.-J. Ihm, M.-S. Yang, and S.-G. Lee, “CMOS low-noise amplifier design optimization techniques,” *IEEE Trans. Microw. Theory Tech.*, vol. 52, no. 5, pp. 1433–1442, 2004.

[15] C. A. Balanis, *Antenna theory: Analysis and Design.* Wiley, 2005, vol. 1.

[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* Wiley, 2012.

[17] M. Loy, “Understanding and enhancing sensitivity in receivers for wireless applications,” *Texas Instruments, Technical Brief*, 1999.

[18] I. E. Telatar, “Capacity of multi-antenna Gaussian channels,” *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, 1999.

[19] P. Almers, E. Bonek, A. Burr, N. Czink, M. Debbah, V. Degli-Esposti, H. Hofstetter, P. Kyösti, D. Laurenson, G. Matz, A. Molisch, C. Oestges, and H. Özcelik, “Survey of channel and radio propagation models for wireless MIMO systems,” *EURASIP J. Wireless Commun. Networking*, no. 1, p. 019070, Feb 2007.

[20] M. L. Morris and M. Jensen, “Network model for MIMO systems with coupled antennas and noisy amplifiers,” *IEEE Trans. Antennas Propag.*, vol. 53, no. 1, pp. 545–552, 2005.

[21] C. Oestges and B. Clerckx, *MIMO Wireless Communications: From Real-world Propagation to Space-Time Code Design.* Academic Press, 2010.

[22] A. Molisch, *Wireless Communications.* Wiley-IEEE Press, 2005.

[23] R. Horn and C. Johnson, *Matrix Analysis.* Cambridge, 1985.

[24] D. Nie, B. M. Hochwald, and E. Stauffer, “Systematic design of large-scale multiport decoupling networks,” *IEEE Trans. Circuits Syst. I*, vol. 61, no. 7, 2014.

[25] A. Voors, “4nec2, NEC based antenna modeler and optimizer,” 2015, [accessed 4-October-2015]. [Online]. Available: http://www.qsl.net/4nec2/

[26] J. W. Wallace and M. Jensen, “Mutual coupling in MIMO wireless systems: a rigorous network theory analysis,” *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, 2004.

[27] R. M. Fano, “Theoretical limitations on the broadband matching of arbitrary impedances,” *J. Franklin Institute*, vol. 249, no. 1, pp. 57–83, 1950.

[28] W.-K. Chen, “Mathematical theory of broadband matching of multiport networks,” *J. Franklin Institute*, vol. 326, no. 5, pp. 737–747, 1989.

[29] P. S. Taluja and B. L. Hughes, “Bandwidth limitations and broadband matching for coupled multi-antenna systems,” in *IEEE Global Telecommun. Conf.*, 2011, pp. 1–6.

[30] L. Kundu and B. L. Hughes, “The impact of frequency-selective matching on the capacity of compact MIMO systems,” in *IEEE Int. Conf. Commun.*, 2014, pp. 2215–2220.

[31] T. Marzetta, "Noncooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[32] H. Q. Ngo, E. Larsson, and T. Marzetta, "Energy and Spectral Efficiency of Very Large Multiuser MIMO Systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[33] C. Desset, B. Debaillie, V. Giannini, A. Fehske, G. Auer, H. Holtkamp, W. Wajda, D. Sabella, F. Richter, M. J. Gonzalez *et al.*, "Flexible power modeling of LTE base stations," in *IEEE Wireless Commun. Networking Conf.*, 2012, pp. 2858–2862.

[34] E. McCune, "Fundamentals for energy-efficient massive MIMO," in *IEEE Wireless Commun. Networking Conf. Workshops*, 2017, pp. 1–6.

[35] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid beamforming for massive MIMO: A survey," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 134–141, 2017.

[36] E. McCune, "Operating modes of dynamic power supply transmitter amplifiers," *IEEE Trans. Microwave Theory Techniq.*, vol. 62, no. 11, pp. 2511–2517, 2014.

[37] J. Zhang, Y. Huang, J. Wang, B. Ottersten, and L. Yang, "Per-antenna constant envelope precoding and antenna subset selection: A geometric approach," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6089–6104, Dec 2016.

[38] H. Yang and T. L. Marzetta, "Performance of conjugate and zero-forcing beamforming in large-scale antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 172–179, February 2013.

[39] G. Reeves, H. D. Pfister, and A. Dytso, "Mutual information as a function of matrix snr for linear gaussian channels," in *IEEE Int. Symp. Inf. Theory*, 2018, pp. 1754–1758.

[40] M. Joham, W. Utschick, and J. A. Nossek, “Linear transmit processing in MIMO communications systems,” *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2700–2712, Aug 2005.

[41] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, “Nonlinear 1-Bit Precoding for Massive MU-MIMO with Higher-Order Modulation,” *ArXiv e-prints*, Dec. 2016.

[42] S. Jacobsson, O. Castañeda, C. Jeon, G. Durisi, and C. Studer, “Nonlinear Phase-Quantized Constant-Envelope Precoding for Massive MU-MIMO-OFDM,” *ArXiv e-prints*, Oct. 2017.

[43] J. H. Winters, “Optimum combining in digital mobile radio with cochannel interference,” *IEEE Trans. Veh. Technol.*, vol. 33, no. 3, pp. 144–155, Aug 1984.

[44] G. Caire and S. Shamai, “On achievable rates in a multi-antenna Gaussian broadcast channel,” in *IEEE Int. Symp. Inf. Theory*, June 2001, p. 147.

[45] P. Viswanath and D. N. C. Tse, “Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality,” *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1912–1921, Aug 2003.

[46] S. Vishwanath, N. Jindal, and A. Goldsmith, “Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels,” *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2658–2668, Oct 2003.

[47] M. Sharif and B. Hassibi, “On the capacity of MIMO broadcast channels with partial side information,” *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 506–522, 2005.

[48] N. Jindal and A. Goldsmith, “Dirty-paper coding versus TDMA for MIMO broadcast channels,” *IEEE Trans. Inf. Theory*, vol. 51, no. 5, pp. 1783–1794, 2005.

[49] I. CVX Research, “CVX: Matlab software for disciplined convex programming, version 2.0,” http://cvxr.com/cvx, Aug. 2012.

[50] M. Grant and S. Boyd, “Graph implementations for nonsmooth convex programs,” in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110.

[51] W. Yu and J. M. Cioffi, “Sum capacity of Gaussian vector broadcast channels,” *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1875–1892, Sept 2004.

[52] T. Yoo and A. Goldsmith, “On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming,” *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 528–541, March 2006.

[53] M. Stojnic, H. Vikalo, and B. Hassibi, “Rate maximization in multi-antenna broadcast channels with linear preprocessing,” *IEEE Trans. Wireless Commun.*, vol. 5, no. 9, pp. 2338–2342, Sept. 2006.

[54] F. Boccardi, F. Tosato, and G. Caire, “Precoding schemes for the MIMO-GBC,” in *Int. Zurich Seminar Commun.*, Feb 2006, pp. 10–13.

[55] A. Wiesel, Y. C. Eldar, and S. Shamai, “Zero-forcing precoding and generalized inverses,” *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 4409–4418, 2008.

[56] M. Joham, W. Utschick, and J. A. Nossek, “Linear transmit processing in MIMO communications systems,” *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2700–2712, 2005.

[57] E. Björnson, M. Bengtsson, and B. Ottersten, “Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure [lecture notes],” *IEEE Signal Process. Mag.*, vol. 31, no. 4, pp. 142–148, July 2014.

[58] G. Caire, “On the ergodic rate lower bounds with applications to massive MIMO,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3258–3268, May 2018.

[59] A. S. Motahari and A. K. Khandani, “Capacity bounds for the Gaussian interference channel,” *IEEE Trans. Inf. Theory*, vol. 55, no. 2, 2009.

[60] X. Shang, G. Kramer, and B. Chen, "A new outer bound and the noisy-interference sum–rate capacity for Gaussian interference channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 689–699, Feb 2009.

[61] V. S. Annapureddy and V. V. Veeravalli, "Gaussian interference networks: Sum capacity in the low-interference regime and new outer bounds on the capacity region," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3032–3050, July 2009.

[62] G. Kaplan and S. Shamai, "Information rates and error exponents of compound channels with application to antipodal signaling in a fading environment," *AEU. Archiv Elektr. Übertrag.*, vol. 47, no. 4, pp. 228–239, 1993.

[63] A. Ganti, A. Lapidoth, and I. E. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2315–2328, Nov. 2000.

[64] G. Böcherer, F. Steiner, and P. Schulte, "Bandwidth efficient and rate-matched low-density parity-check coded modulation," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 4651–4665, Dec 2015.

[65] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein, and C. Studer, "Quantized Precoding for Massive MU-MIMO," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4670–4684, Nov. 2017.

[66] C. Risi, D. Persson, and E. G. Larsson, "Massive MIMO with 1-bit ADC," *arXiv:1404.7736*, Apr. 2014.

[67] C. K. Wen, C. J. Wang, S. Jin, K. K. Wong, and P. Ting, "Bayes-Optimal Joint Channel-and-Data Estimation for Massive MIMO With Low-Precision ADCs," *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2541–2556, May 2016.

[68] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "Throughput Analysis of Massive MIMO Uplink With Low-Resolution ADCs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 4038–4051, Jun. 2017.

[69] A. K. Saxena, I. Fijalkow, and A. L. Swindlehurst, “Analysis of One-Bit Quantized Precoding for the Multiuser Massive MIMO Downlink,” *IEEE Trans. Signal Process.*, vol. 65, no. 17, pp. 4624–4634, Sep. 2017.

[70] H. Jedda, J. A. Nossek, and A. Mezghani, “Minimum BER precoding in 1-bit massive MIMO systems,” in *IEEE Sensor Array Multich. Signal Process. Workshop*, 2016, pp. 1–5.

[71] M. Staudacher, G. Kramer, W. Zirwas, B. Panzner, and R. S. Ganesan, “Optimized combination of conventional and constrained massive MIMO arrays,” in *ITG Workshop Smart Antennas*, March 2017, pp. 1–4.

[72] C.-J. Wang, C.-K. Wen, S. Jin, and S.-H. Tsai, “Finite-Alphabet Precoding for Massive MU-MIMO with Low-resolution DACs,” *arXiv:1709.05755*, Sep. 2017.

[73] S. Jacobsson, G. Durisi, M. Coldrey, and C. Studer, “Linear Precoding with Low-Resolution DACs for Massive MU-MIMO-OFDM Downlink,” *arXiv:1709.04846*, Sep. 2017.

[74] H. Jedda, A. Mezghani, J. A. Nossek, and A. L. Swindlehurst, “Massive MIMO downlink 1-bit precoding with linear programming for PSK signaling,” in *IEEE Int. Workshop Signal Process. Advances Wireless Commun.*, July 2017, pp. 1–5.

[75] F. Askerbeyli, H. Jedda, and J. A. Nossek, “1-Bit Precoding in Massive MU-MISO-OFDM Downlink with Linear Programming,” in *Int. ITG Workshop Smart Antennas (WSA)*, Vienna, April 2019, pp. 1–5.

[76] A. Nedelcu, F. Steiner, M. Staudacher, G. Kramer, W. Zirwas, R. S. Ganesan, P. Baracca, and S. Wesemann, “Quantized precoding for multi-antenna downlink channels with MAGIQ,” in *Int. VDE ITG Workshop on Smart Antennas*, Mar. 2018, pp. 1–8.

[77] C. G. Tsinos, A. Kalantari, S. Chatzinotas, and B. Ottersten, “Symbol-level precoding with low resolution DACs for large-scale array MU-MIMO systems,” in *IEEE Int. Workshop Signal Process. Advances Wireless Commun.*, Jun. 2018, pp. 1–5.

[78] S. Domouchtsidis, C. G. Tsinos, S. Chatzinotas, and B. Ottersten, "Symbol-level precoding for low complexity transmitter architectures in large-scale antenna array systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 852–863, 2019.

[79] M. A. Sedaghat, A. Bereyhi, and R. R. Müller, "Least square error precoders for massive MIMO with signal constraints: Fundamental limits," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 667–679, 2017.

[80] G. Parisi, M. Mézard, and M. A. Virasoro, "Spin glass theory and beyond," *World Scientific, Singapore*, vol. 187, p. 202, 1987.

[81] H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering.* Springer Publishing Company, Incorporated, 2013.

[82] Y. Liu, Y. Dai, and Z. Luo, "Coordinated beamforming for MISO interference channel: Complexity analysis and efficient algorithms," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 1142–1157, March 2011.

[83] O. Toker and H. Ozbay, "On the complexity of purely complex SPL mu computation and related problems in multidimensional systems," *IEEE Trans. Autom. Control*, vol. 43, no. 3, pp. 409–414, March 1998.

[84] T. Tong Wu and K. Lange, "Coordinate descent algorithms for lasso penalized regression," *ArXiv e-prints*, Mar. 2008.

[85] S. C. Dafermos and F. T. Sparrow, "The traffic assignment problem for a general network," *J. Research Nat. Bureau of Standards B*, vol. 73, no. 2, pp. 91–118, 1969.

[86] A. Beck, *First-Order Methods in Optimization.* Philadelphia, PA: Society for Industrial and Applied Mathematics, 2017. [Online]. Available: https://epubs.siam.org/doi/abs/10.1137/1.9781611974997

[87] J. Nutini, M. Schmidt, I. H. Laradji, M. Friedlander, and H. Koepke, "Coordinate descent converges faster with the Gauss-Southwell rule than random selection," in *Int. Conf. Machine Learning*, ser. ICML'15, 2015, pp. 1632–1641.

[88] D. Gesbert, T. Ekman, and N. Christophersen, "Capacity limits of dense palm-sized MIMO arrays," in *IEEE Global Telecommun. Conf.*, vol. 2, Nov 2002, pp. 1187–1191 vol.2.

[89] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge, 2004.

[90] M. X. Goemans and D. P. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming," *J. ACM*, vol. 42, no. 6, pp. 1115–1145, Nov. 1995.

[91] Y. Nesterov, "Semidefinite relaxation and nonconvex quadratic optimization," *Optimization Methods and Software*, vol. 9, no. 1-3, pp. 141–160, 1998.

[92] A. M. So, "Probabilistic analysis of the semidefinite relaxation detector in digital communications," in *ACM-SIAM Symp. Discrete Algorithms*, 2010, pp. 698–711.

[93] W.-K. Ma, P.-C. Ching, and Z. Ding, "Semidefinite relaxation based multiuser detection for M-ary PSK multiuser systems," *IEEE Trans. Signal Process.*, vol. 52, no. 10, pp. 2862–2872, Oct 2004.

[94] X. Fan, J. Song, D. P. Palomar, and O. C. Au, "Universal binary semidefinite relaxation for ML signal detection," *IEEE Trans. Commun.*, vol. 61, no. 11, pp. 4565–4576, November 2013.

[95] C. Lu, Y. Liu, and J. Zhou, "An efficient global algorithm for nonconvex complex quadratic problems with applications in wireless communications," in *IEEE-CIC Int. Conf. Commun. China*, Oct 2017, pp. 1–5.

[96] C. Lu, Z. Deng, W.-Q. Zhang, and S.-C. Fang, "Argument division based branch-and-bound algorithm for unit-modulus constrained complex quadratic programming," *J. Global Optim.*, vol. 70, no. 1, pp. 171–187, Jan 2018.

[97] C. Buchheim and A. Wiegele, “Semidefinite relaxations for non-convex quadratic mixed-integer programming,” *Math. Progr.*, vol. 141, no. 1, pp. 435–452, Oct 2013.

[98] U. Fincke and M. Pohst, “Improved methods for calculating vectors of short length in a lattice, including a complexity analysis,” *Math. Comp.*, vol. 44, no. 170, pp. 463–471, 1985.

[99] J. Jaldén and B. Ottersten, “On the complexity of sphere decoding in digital communications,” *IEEE Trans. Signal Process.*, vol. 53, no. 4, pp. 1474–1484, 2005.

[100] Z. Guo and P. Nilsson, “Algorithm and implementation of the K-best sphere decoding for MIMO detection,” *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 491–503, 2006.

[101] S. Yang and L. Hanzo, “Fifty years of MIMO detection: The road to large-scale MIMOs,” *IEEE Commun. Surveys & Tut.*, vol. 17, no. 4, pp. 1941–1988, 2015.

[102] S. K. Mohammed and E. G. Larsson, “Per-antenna constant envelope precoding for large multi-user MIMO systems,” *IEEE Trans. Commun.*, vol. 61, no. 3, pp. 1059–1071, March 2013.

[103] F. A. Dietrich, R. Hunger, M. Joham, and W. Utschick, “Robust transmit Wiener filter for time division duplex systems,” in *IEEE Int. Symp. Signal Proc. and Inf. Tech.*, Dec 2003, pp. 415–418.

[104] A. Gleixner, L. Eifler, T. Gally, G. Gamrath, P. Gemander, R. L. Gottwald, G. Hendel, C. Hojny, T. Koch, M. Miltenberger *et al.*, “The SCIP optimization suite 5.0,” https://opus4.kobv.de/opus4-zib/frontdoor/index/index/docId/6629f, 2017.

[105] P. Kyösti, “IST-4-027756 WINNER II D1.1.2 V1.2,” https://www.cept.org/files/8339/winner2%20-%20final%20report.pdf, 2007, [Online; accessed May 2020].

[106] H. Pirzadeh, G. Seco-Granados, A. L. Swindlehurst, and J. A. Nossek, “On the effect of mutual coupling in one-bit spatial sigma-delta massive MIMO systems,” in *IEEE Int. Workshop Signal Proc. Advances Wireless Commun.*, 2020, pp. 1–5.

[107] B. K. Lau and J. B. Andersen, “Simple and efficient decoupling of compact arrays with parasitic scatterers,” *IEEE Trans. Antennas Propag.*, vol. 60, no. 2, pp. 464–472, 2011.

[108] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.

[109] R. G. Gallager, *Information Theory and Reliable Communication.* John Wiley & Sons, Inc., 1968.

[110] D. M. Arnold, H. A. Loeliger, P. O. Vontobel, A. Kavcic, and W. Zeng, “Simulation-Based Computation of Information Rates for Channels With Memory,” *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3498–3508, Aug. 2006.

[111] M. X. Goemans and D. P. Williamson, “Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming,” *J. ACM*, vol. 42, no. 6, pp. 1115–1145, 1995.

[112] D. Micciancio, “Inapproximability of the shortest vector problem: Toward a deterministic reduction,” *Theory of Computing*, vol. 8, no. 1, pp. 487–512, 2012.

[113] A. d’Aspremont and S. Boyd, “Relaxations and randomized methods for nonconvex QCQPs,” https://see.stanford.edu/materials/lsocoee364b/AdditionalLecture1-relaxations.pdf, 2003.

[114] C. Buchheim, M. Montenegro, and A. Wiegele, “SDP-based branch-and-bound for non-convex quadratic integer optimization,” *J. Global Optim.*, vol. 73, no. 3, pp. 485–514, 2019.