Fachgebiet Methoden der Signalverarbeitung
Technische Universität München

Subarray Hybrid Architectures for mmWave Communication

Marcin Iwanow

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

Doktor-Ingenieurs

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. -Ing. habil. Erwin Biebl
Prüfer der Dissertation:	1. Prof. Dr. -Ing. Wolfgang Utschick
	2. Prof. Luis Castedo Ribas, Ph.D.
	3. -

Die Dissertation wurde am 25.09.2020 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 28.04.2021 angenommen.

Contents

Symbols

N_{r}	Number of receive antennas
$N_{\mathrm{r}}^{\mathrm{RF}}$	Number of receiver RF chains
N_{r}^{l}	Number of antennas at the lth receive subarray
$N_{\mathrm{r}}^{\mathrm{RF},l}$	Number of RF chains at the lth receive subarray
$\boldsymbol{G}$	Combiner matrix at the Rx
$\boldsymbol{G}_A$	Analog combiner matrix at the Rx
$\boldsymbol{G}_D$	Digital combiner matrix at the Rx
$\boldsymbol{G}_A^{(l)}$	Analog combiner matrix at the lth receive subarray
$\boldsymbol{G}_D^{(k)}$	Digital combiner matrix at the lth receive subarray
$\xi_{A,\mathrm{RX}}^{(l)}$	Feasible set for the analog combiner matrices at the Rx
$\boldsymbol{G}^{(l)}$	Combiner matrix at the lth receive subarray
$\boldsymbol{y}$	Received signal vector
N_{t}	Number of the transmit antennas
$N_{\mathrm{t}}^{\mathrm{RF}}$	Number of the transmitter RF chains
N_{t}^{k}	Number of antennas at the kth transmit subarray
$N_{\mathrm{t}}^{\mathrm{RF},k}$	Number of the RF chains at the kth transmit subarray
S_{t}	Number of transmit subarrays
$\boldsymbol{P}$	Precoder matrix at the Tx
$\boldsymbol{P}_A$	Analog precoder matrix at the Tx
$\boldsymbol{P}_D$	Digital precoder matrix at the Tx
$\boldsymbol{P}_A^{(k)}$	Analog precoder matrix at the kth subarray at the Tx
$\boldsymbol{P}_D^{(k)}$	Digital precoder matrix at the kth subarray at the Tx

$\xi_{A,\mathrm{TX}}^{(k)}$	Feasible set for the analog precoder matrices at the Tx
$\boldsymbol{P}^{(k)}$	Precoder matrix at the kth subarray at the Tx
$\boldsymbol{P}_{\mathrm{FD}}^{\star}$	Optimal precoder matrix at the Tx
$\boldsymbol{x}$	Vector transmitted from the Tx antennas
$\boldsymbol{R}_x$	Covariance matrix of $\boldsymbol{x}$
N_{s}	Number of streams
N_{subc}	Number of subcarriers
P_{C}	Power consumed by the transmitter
P_{E}	Power emitted by the transmitter
$\boldsymbol{\eta}$	Receiver noise vector
$\boldsymbol{R}_\eta$	Covariance matrix of $\boldsymbol{\eta}$
$\boldsymbol{H}$	Channel matrix
$\boldsymbol{H}_{l,k}$	Channel matrix between the kth Tx subarray and the lth Rx subarray.
$\boldsymbol{s}$	Vector of symbols

Notation

$\mathbb{R}$	Set of real numbers
$\mathbb{R}_+$	Set of positive real numbers including 0
$\mathbb{C}$	Set of complex numbers
$\mathbb{C}^N$	Set of column vectors with N complex entries
$\mathbb{C}^{N\times M}$	Set of matrices with N rows and M columns with complex entries
$\mathrm{diag}(\bullet)$	Vector containing entries lying on the diagonal of the matrix
$\mathrm{vec}(\bullet)$	Vectorization of a matrix
$\mathrm{rank}(\bullet)$	Matrix rank
$\det(\bullet)$	Matrix determinant
$(\bullet)^{\mathrm{T}}$	Transpose of a matrix
$(\bullet)^{\mathrm{H}}$	Conjugate transpose of a matrix
$(\bullet)^{-1}$	Inverse of a matrix
$(\bullet)^{\dagger}$	Moore-Penrose pseudoinverse of a matrix
$\Vert \bullet \Vert_p$	p-norm of a vector/matrix
$\Vert \bullet \Vert_{\mathrm{F}}$	Frobenius norm of a matrix
$\mathrm{E}[\bullet]$	Expectation of a random variable
$p_x(\bullet)$	Probability density function of the random variable x
$I(x, y)$	Mutual information between random variables x and y
$\mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu}, \boldsymbol{C})$	Circularly symmetric multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{C}$

$\mathcal{N}_{\mathbb{C}}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{C})$	Probability density function of a random vector $\boldsymbol{x}$, that follows the circularly symmetric multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{C}$
$(\bullet)_{r_1:r_2,c_1:c_2}$	Submatrix of the matrix in the argument, containing rows in the range of r_1 to r_2 and columns in the range of c_1 to c_2. If $r_1 = r_2$, we replace $r_1 : r_2$ by r_1 and we do analogously for the column indices.

Abbreviations

i.i.d	Identical and independently distributed
FD	Full digital
FER	Frame error rate
MF	Matched filter
ZF	Zero forcing
PSN	Phase shifter network
RF	Radio frequency
Rx	Receiver
Tx	Transmitter
SGSHP	Spatial Greedy Subarray Hybrid Precoding
MIMO	Multiple-input multiple-output
mmWave	Millimeter-wave
5G	Fifth-generation mobile networks
DAC	Digital-Analog converter
ADC	Analog-digital converter
HBF	Hybrid beamforming
FI-HBF	Fully-interconnected hybrid architecture
PI-HBF	Partially-interconnected hybrid architecture
SPI-HBF	Subarray partially-interconnected hybrid architecture
SNR	Signal to Noise Ratio
MAC	Multiple access channel
MSE	Mean square error
BER	Bit error rate
pdf	Probability density function
PAN	Personal Area Network
SotA	State of the Art
DoF	Degrees of freedom

Abstract

Power consumption is a significant obstacle in wide deployment of systems with a large number of antennas. Consequently, it is also a limiting factor for introducing mmWave technology, where large number of antennas is a necessity. Among all ideas to reduce power consumption, the one perhaps most covered in the literature is the hybrid analog-digital architecture of the transceivers. Such approach allows for a significant reduction of power consumption without a substantial sacrifice of the data rates.

However, the widely used version of the hybrid analog-digital transceiver, where the analog circuit is constructed as a phase shifter network connecting all the RF chains and antennas, is not realizable in practice. A practical simplification of the architecture is the subarray partially interconnected analog-digital architecture. There, the phase shifter network is divided into non-overlapping blocks, where in each block a subset of RF chains is interconnected with a subset of antennas.

This work covers a significant part of the analysis of a point-to-point subarray hybrid architecture. Upper bounds for the achievable rate in low- and high-SNR regions are derived. Moreover, we show why the comparisons between the basic and subarray architecture in the literature are often not fair and propose how to handle such a comparison. Lastly, we show how beamforming algorithms can be designed for the subarray architecture by leveraging the ideas invented for the basic fully-interconnected architecture and adopting a new perspective which is unique for the subarray architecture.

1

Introduction

1.1. Thesis overview

Transmission in the $30-300$ GHz frequency range dates back to the very beginnings of wireless systems in the 19th century. However, only recently it has gained a lot of momentum as the technological advances enabled massive deployment of mmWave systems. For example, the WirelessHD standard (aimed on wireless video streaming) proposed transmission in the unlicensed 60 GHz band [3]. Moreover, the IEEE 802.11ad standard specified a Wireless LAN system on mmWave frequencies [2]. For wireless communication systems, it took the publication of [59] to convince the community about the feasibility of the technology for such application. While signal processing techniques which allow for more efficient utilization of the traditional sub-6 GHz frequency bands are constantly developed, the promise of utilizing large portions of spectrum available in mmWave frequencies is definitely tempting. The aforementioned WLAN and PAN standards utilize 2 GHz bandwidths and similar is expected for the cellular communication.

The previously mentioned work from [59] sparked a lot of scientific activity which is supposed to prepare the ground for successful deployment of mmWave in one of the upcoming generations of cellular standards. The activity spans a wide area, i.a., channel modeling activities [5, 27, 58, 62, 63], hardware design [20, 30, 77], identifying MAC layer challenges [10,13,16,23,42], and signal processing solutions for the physical layer [4, 12, 18, 67, 82]. The scope of the thesis lies in the latter category.

In this work, a particular hardware structure which is a possible candidate for

implementation in mmWave transceivers is considered and theoretical investigation of its performance is conducted. The subarray partially interconnected architecture (SPI-HBF), which is the structure in hand, exhibits pronounced hardware simplifications with respect to the more complicated (and widely considered in the literature) fully interconnected hybrid beamforming architecture (FI-HBF). The main contributions in the thesis include:

- theoretical upper bounds for SPI-HBF achievable rate in the low- and high-SNR region,
- extension of the FI-HBF precoding algorithms to SPI-HBF,
- novel precoding strategies for systems with SPI-HBF,
- analysis how using the consumed power constraint (instead of the usual choice of the emitted power) influences the performance comparison of SPI-HBF and FI-HBF along with arguing why such approach is more fair.

The objective of the thesis is to broaden the understanding of the performance of SPI-HBF. Therefore, the idealistic assumptions of full CSI knowledge at the Tx as well as ideal operation of hardware elements, are made.

1.2. Thesis structure

The thesis is organized as follows.

In Chapter 2, we start with introducing the basic concepts of communicating with multiple antennas. We explain the gains that we experience with such communication. Further, we discuss the capacity of MIMO systems and performance measures for practical MIMO systems. Finally, we explain the significance of precoding and combining in MIMO. Moreover, we describe the specifics of communication on mmWave frequencies. We discuss the propagation characteristics and the influence on the communication channel thereof. We justify the need of massive antenna arrays at the mmWave transceivers and present challenges of such requirement. Further, we discuss how the challenges can be overcome and as a consequence, communication on mmWave frequency range can be enabled.

In Chapter 3, we explain the system model which we consider in the thesis. We formulate the input-output relation of the system and discuss its specifics due to the

operating frequency band. Also, we introduce the performance measure used for the system. Furthermore, we discuss how the analog stage of SPI-HBF can be modeled mathematically.

Before proposing practical precoding schemes for SPI-HBF, we show in Chapter 4 that it is possible to derive tighter upper bounds for the achievable rates at low and high SNR than the ones usually used in the literature. Deriving a tight upper bound is very important in the process of designing practical algorithms—knowing that the performance is close to the limit can focus the efforts on other issues like complexity etc.

In Chapter 5, we present the most common approaches to precoding when HBF architectures are considered. We show novel, optimal ways of optimizing the digital precoder in schemes that rely on alternating optimization methods for minimizing the difference of hybrid precoding and optimal unconstrained precoding. Furthermore, we propose a greedy precoding scheme which is especially suited for SPI-HBF and is less complex than other algorithms. Also, we propose a precoding algorithm which is specifically designed for scenarios with low-SNR.

We finish with a summary of the results in Chapter 6 and an outlook for further work on the topic.

2

Communication with Multiple Antennas

The development of digital communication dates back to Shannon [64]. There it has been shown that transmission with presence of noise may be reliable—if the transmission rate doesn't exceed capacity, the error can be made arbitrarily small with increasing length of the transmission block. A similar breakthrough for multi-antenna communication has been marked by the Telatar's and Foschini's works on MIMO channel capacity [17, 71] and Alamouti's publication on transmit diversity with MIMO [6].

In this chapter, we start with introducing the basic concepts of communicating with multiple antennas. We explain the gains that we experience with such communication. Further, we discuss the capacity of MIMO systems and performance measures for practical MIMO systems. Finally, we explain the significance of precoding and combining in MIMO.

2.1. Introduction to MIMO

A point-to-point MIMO system consists of a transmitter with $N_{\mathrm{t}} \geq 1$ transmit antennas and a receiver with $N_{\mathrm{r}} \geq 1$ receive antennas. In a narrowband system (when the product of the delay spread and the bandwidth is small) the propagation between each transmit antenna k and receive antenna l is characterized by a complex number $h_{l,k}$, called the channel coefficient. Thus, the propagation between all antenna pairs

in the system may be collected in a matrix

$$\boldsymbol{H} = \begin{bmatrix} h_{1,1} & \dots & h_{1,N_\mathrm{t}} \\ \vdots & \ddots & \vdots \\ h_{N_\mathrm{r},1} & \dots & h_{N_\mathrm{r},N_\mathrm{t}} \end{bmatrix} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{t}} \tag{2.1}$$

called further the channel matrix. Intuitively, it should be possible to benefit from having at disposal the entire channel matrix $\boldsymbol{H}$ instead of the single channel coefficient h like in SISO. This is indeed the case. More specifically, MIMO can be utilized in order to improve reliability and achievable rates of the link.

In order to improve reliability, the same information can be transmitted between all the $N_\mathrm{t}N_\mathrm{r}$ antenna pairs. This may result in up to $d = N_\mathrm{t}N_\mathrm{r}$ diversity order of the link (given statistical independence of the channel coefficients between different antenna pairs) and the error rate decaying like SNR^{-d} as compared to SNR^{-1} in a SISO link [87].

The increase in achievable rates can be achieved by utilizing spatial degrees of freedom. A MIMO channel may be decomposed into $\min\{N_\mathrm{t}, N_\mathrm{r}\}$ parallel channels (given full column- or row-rank of the channel matrix $\boldsymbol{H}$ influenced by, e.g., antenna spacing, scattering etc.) resulting in a linear increase in capacity at high-SNR. Independent transmission of multiple data streams through the MIMO channel is referred to as multiplexing.

Finally, a trade-off between the two gains is possible, leading to the fundamental diversity-multiplexing trade-off in MIMO [87]. In general, exploiting the diversity order of the channel is crucial at low-SNR, when the system is power limited and exploiting the multiplexing gain is important in the interference limited high-SNR region.

In the point-to-multipoint scenario the significance of MIMO transceivers becomes even more pronounced. In addition to improving the reliability and rates within the system, multiple antennas have become an important tool in managing the interference between users transmitting at the same time and frequency resources.

2.2. Capacity in a MIMO system

Let's assume that $x_k \in \mathbb{C}$ is fed to the kth transmit antenna and $y_l \in \mathbb{C}$ is received by the lth receive antenna. Assuming that signals from all the transmit antennas are

superimposed ideally at each receive antenna, we write

$$y_l = \sum_{k=1}^{N_\mathrm{t}} h_{l,k} x_k + \eta_l. \tag{2.2}$$

After grouping the transmitted and received signals into vectors

$$\boldsymbol{x} = [x_1, \ldots, x_{N_\mathrm{t}}]$$
$$\boldsymbol{y} = [y_1, \ldots, y_{N_\mathrm{r}}]$$

and assuming additive noise grouped in a vector $\boldsymbol{\eta} \in \mathbb{C}^{N_\mathrm{r}}$ at the receiver antennas, the input-output relation is linear and reads

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{\eta}. \tag{2.3}$$

Shannon showed that for a time-invariant channel $\boldsymbol{H}$, the capacity is equal to the mutual information between $\boldsymbol{x}$ and $\boldsymbol{y}$ maximized over all the possible distributions of $\boldsymbol{x}$

$$C(\boldsymbol{H}) = \max_{p(\boldsymbol{x})} I(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{H}). \tag{2.4}$$

Similar notions of capacity can be defined if [17, 24, 71, 74]

- the channel is time-variant but can be tracked and the transmit strategy may be adapted for each channel realization or
- the channel is rapidly changing such that during transmission of each codeword every possible channel state is experienced; the transmit strategy does not adapt to the instantaneous channel. The channel is memoryless.

The *ergodic* capacity is defined for the first case as

$$C_{e,1} = \mathrm{E}_{\boldsymbol{H}} \left[\max_{p(\boldsymbol{x})} I(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{H}) \right] = \mathrm{E}_{\boldsymbol{H}} \left[C(\boldsymbol{H}) \right] \tag{2.5}$$

and for the second case as

$$C_{e,2} = \max_{p(\boldsymbol{x})} \mathrm{E}_{\boldsymbol{H}} \left[I(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{H}) \right]. \tag{2.6}$$

The first expression is essentially the maximum mutual information averaged over the distribution of the random variable $\boldsymbol{H}$. In the second expression, the maximization

of the (average) mutual information is justified as the channel is memoryless, i.e., each channel realization is independent.

However, this holds no longer for very slow varying channels. Then, each codeword experiences a different subset of possible realizations of the channels (the process is no longer ergodic). Therefore, the probability that the rate falls below any possible threshold is nonzero, which leads to the Shannon capacity equal to zero. For this class of channels the notion of *outage capacity* is defined as

$$C_{\text{out}}(\rho) = \max R \quad \text{s.t.} \;\; p_{\boldsymbol{H}}(R > C(\boldsymbol{H})) \leq \rho \tag{2.7}$$

which is the maximum rate for which the probability of outage (transmission with a rate exceeding capacity) falls below a given threshold ρ.

In MIMO systems with many users the capacity region is a set of tuples corresponding to all possible rates that can be simultaneously achieved by all users in the system [24].

2.3. Perfomance Measures

There does not exist a universal performance measure for communication links. The three most common ones are the achievable rate, frame/bit error ratio [52], and the mean square error (MSE). While the last one is the least natural, it's relationship with the achievable rate makes it a relevant metric.

For a given transmit strategy, i.e., a given distribution $p'(\boldsymbol{x})$ of the input vector $\boldsymbol{x}$,

- the achievable rate is the mutual information between $\boldsymbol{x}$ and $\boldsymbol{y}$ with $\boldsymbol{H}$ being a parameter

$$R = I(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{H})_{|\boldsymbol{x}\sim p'(\boldsymbol{x})}, \tag{2.8}$$

- the MSE is the averaged squared norm of the error between the estimate $\hat{\boldsymbol{x}}$ of the transmit vector $\boldsymbol{x}$ and the actual value thereof

$$\text{MSE} = \text{E}_{\boldsymbol{H}} \left[\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2 \right]_{|\boldsymbol{x}\sim p'(\boldsymbol{x})}, \tag{2.9}$$

- the BER/FER quantifies the ratio of bits/frames which were erroneously decoded.

Apart from these, there exist a large number of performance measures related to specific requirements of different systems.

2.4. Precoding and Combining

The usual building blocks of a SISO transmitter consist of the channel/source coder, interleaver, and symbol mapper. This must be complemented in MIMO transmitters with two additional blocks—the space-time encoder and space-time precoder [14]. The space-time encoder maps the stream of modulated symbols to N_s spatial data streams. The space-time precoder distributes the spatial data streams among the antennas. We denote the output of the space-time encoder—further referred to as the data vector—with $\boldsymbol{s} \in \mathbb{C}^{N_\mathrm{s}}$.

Precoding is a crucial operation in MIMO systems. It represents the strategy for using the additional, spatial degree of freedom in MIMO. The designer of the system may optimize the precoding strategy based on the system requirements—for example minimize the MSE of the data vector reconstruction, impose conditional unbiasedness of the data vector reconstruction (a.k.a. zero forcing), or maximize the achievable data rate.

A large portion of literature analyzes linear precoding strategies (such that $\boldsymbol{x} = \boldsymbol{P}\boldsymbol{s}$). A comprehensive study thereof is presented in [39]. This is not only due to reduced complexity with respect to non-linear approaches, but also due to near optimal or even optimal performance in many crucial cases. For example, if the noise at the receiver follows circularly symmetric complex Gaussian distribution $\boldsymbol{\eta} \sim \mathcal{N}_\mathbb{C}(\boldsymbol{0}, \boldsymbol{R}_\eta)$ and the channel matrix $\boldsymbol{H}$ is known to the transmitter, the capacity-achieving distribution of $\boldsymbol{x}$ is also circular symmetric complex Gaussian with $\boldsymbol{0}$ mean and covariance matrix $\boldsymbol{R}_x{}^\star$ maximizing the achievable rate [71], i.e.,

$$\boldsymbol{R}_x{}^\star = \arg\max_{\boldsymbol{R}_x \succeq 0} R \tag{2.10}$$

with

$$R = I(\boldsymbol{x}, \boldsymbol{y}; H)_{|\boldsymbol{x} \sim \mathcal{N}_\mathbb{C}(\boldsymbol{0}, \boldsymbol{R}_x)} = \log_2 \det\left(\boldsymbol{I} + \boldsymbol{R}_\eta^{-1} \boldsymbol{H} \boldsymbol{R}_x \boldsymbol{H}^\mathrm{H}\right). \tag{2.11}$$

Then, given the data vector is an i.i.d. circular symmetric complex Gaussian variable $\boldsymbol{s} \sim \mathcal{N}_\mathbb{C}(\boldsymbol{0}, \boldsymbol{I})$, the optimal precoding is a linear function $\boldsymbol{x} = \boldsymbol{P}^\star \boldsymbol{s}$ such that $\boldsymbol{P}^\star \boldsymbol{P}^{\star\mathrm{H}} = \boldsymbol{R}_x{}^\star$, a decomposition that is possible for any covariance matrix.

The combining operation (often referred to as equalization) is the counterpart of the precoding at the receiver. It combines the signal $\boldsymbol{y}$ received at the antennas into $\hat{\boldsymbol{s}} \in \mathbb{C}^{N_\mathrm{s}}$, which is the reconstruction of the data vector $\boldsymbol{s}$. Similarly as the precoder, it may be constructed optimally w.r.t. a chosen performance measure.

Interestingly, the achievable rate of the link is usually maximized by a large set of combiners. It is required to preserve in $\hat{\boldsymbol{s}}$ sufficient statistics of $\boldsymbol{s}$ included in $\boldsymbol{y}$. In the case of the Gaussian channel described earlier, the combining function that maximizes the achievable rate is linear with the form $\boldsymbol{X}\boldsymbol{P}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{R}_{\eta}^{-1}$ with $\boldsymbol{X}$ being any invertible matrix. However, while all such combiners maximize the achievable rate, they may exhibit different properties. For example, while three most common combiners—MMSE (Wiener), ZF, and MF are of the required form with[1]

$$\begin{aligned}\boldsymbol{X}_{\mathrm{MF}} &= \boldsymbol{I}\\ \boldsymbol{X}_{\mathrm{ZF}} &= \left(\boldsymbol{P}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}{\boldsymbol{R}_{\eta}}^{-1}\boldsymbol{H}\boldsymbol{P}\right)^{-1}\\ \boldsymbol{X}_{\mathrm{MMSE}} &= \left(\boldsymbol{I}+\boldsymbol{P}^{\mathrm{H}}\boldsymbol{H}^{\mathrm{H}}{\boldsymbol{R}_{\eta}}^{-1}\boldsymbol{H}\boldsymbol{P}\right)^{-1},\end{aligned}$$

they have different properties in terms of the error rates [39].

[1]We note that the identity matrix in the matched filter and MMSE filter expressions stems from the assumption that the covariance matrix of $\boldsymbol{s}$ is an identity matrix. Otherwise, the identity matrix should be replaced by the inverse of the covariance matrix of $\boldsymbol{s}$.

3

Communication on mmWave frequencies

3.1. Propagation on mmWave Frequencies

The key to system and algorithm design for mmWave lies in understanding how the propagation characteristics differ with respect to the lower frequencies. There has been significant work on this topic which has been summarized, e.g., in [49]. It has been widely agreed that the propagation is dominated by reflections, with diffraction playing a negligible role. Moreover, the number of reflections of a path entails increased propagation distance and, consequently, decrease of the path's power. Therefore, the propagation is defined by a limited number of paths which experience low number of reflections. This results in both significant angular selectivity of the mmWave channel and relatively low delay spread.

Moreover, a common argument that is used for presenting unfavorable propagation characteristics in the mmWave frequency range is the path loss. However, this should be taken with a grain of salt, which we explain in the following. The Friis formula for path gain reads

$$G = G_{\mathrm{Rx}} G_{\mathrm{Tx}} \left(\frac{c}{4\pi d f} \right)^2$$

where $G_{\mathrm{Rx}}/G_{\mathrm{Tx}}$ is the gain of the Rx/Tx antenna (array), c is the speed of light, f is the frequency. This suggests a large pathloss for high frequencies, given the antenna

gains are the same. However, the aperture of antennas with given gain also decreases quadratically with frequency. Consequently, if both the Tx and Rx antennas have the same aperture, the path gain actually increases with frequency.

The propagation characteristics influence both the required hardware and the design of algorithms for mmWave.

First, the number of antennas must be significantly higher than on lower frequencies in order to maintain the antenna aperture, and in consequence the path gains. A large number of antennas comes with many challenges from the hardware perspective. The usual design of transceivers would cause large power consumption—this has been perceived as a problem already in Massive-MIMO (which involves lower frequencies) and is even more pronounced on mmWave frequencies due to higher bandwidths of the signal.

Second, the propagation channel includes only a limited number of individual paths—the usual assumption of rich scattering does not hold any more. Moreover, as the channel is selective in the angular domain, beamforming emerges as a necessity. The high dimensionality of the channel poses also problems in the channel estimation phase.

In the thesis, we address both challenges. First, we discuss different transceiver architectures which are eligible as candidates for the mmWave band. Then, we concentrate on one of them (the subarray hybrid architecture) and perform a thorough analysis thereof—we discuss the theoretical performance limits and present novel algorithmic approaches which outperform the State-of-the-Art (SotA).

3.2. Practical Issues With Large Antenna Arrays

The narrowband fully digital (FD) MIMO system model is often described with the following equation

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{\eta} \tag{3.1}$$

where $\boldsymbol{x} \in \mathbb{C}^{N_\mathrm{t}}$ is a vector of N_t complex inputs to digital-to-analog converters (DACs) at the Tx and, similarly, $\boldsymbol{y} \in \mathbb{C}^{N_\mathrm{r}}$ represents N_r complex outputs from the analog-to-digital converters (ADCs) at the Rx[1]. The noise is denoted with $\boldsymbol{\eta} \in \mathbb{C}^{N_\mathrm{r}}$ and $\boldsymbol{H} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{t}}$ is the channel transfer matrix.

[1]More specifically, the Tx is equipped with $2N_\mathrm{t}$ DACs where the inputs are the real and complex parts of $\boldsymbol{x}$, similarly the Rx is equipped with $2N_\mathrm{r}$ ADCs.

This representation requires taking a number of assumptions about the operation of ADCs/DACs, among others

- infinite resolution,
- sampling with Nyquist rate,
- bandwidth larger than the signal bandwidth,
- availability of 2 ADCs/DACs for each Tx/Rx antenna.

The model in (3.1) holds approximately for high resolution ($12-15$ bit) ADCs/DACs. However, high-resolution converters fulfilling design parameters required for communication on mmWave frequencies (i.a., bandwidth $>$ 500 MHz, sampling rate $>$ 1 GHz) become power hungry, especially the ADCs. The most notable ADC architectures meeting the requirements are the folded-flash, pipeline, and time-interleaved successive approximation. While the folded-flash architecture is the most power hungry (the power dissipation increases exponentially with the resolution), the power consumption of the two others is also significant [41, 79].

The main trends in current research limits the overall power dissipation by either decreasing the resolution of the converters [45, 47, 48, 77], or reducing their number. In the following, we concentrate on the latter.

3.3. Hybrid Transceiver Structure

In order to reduce the number of AD/DA converters, an analog circuit that maps low-dimensional signals at the outputs of the ADCs to high-dimension signals at the inputs of the antennas (and vice versa) is required (c.f. Fig. 3.1). There have been various concepts of such circuits, among them the most widely considered are the lens antenna structure [20, 83, 84] and the phase shifter network (PSN) [7, 8, 12, 22, 30, 35–38, 44, 53, 67, 68, 75–77, 82].

The concept behind the lens antenna is to introduce a fixed time delay network between N antennas and K ports, where K must be greater than the number of RF chains N_{RF} (where the RF chains consist, i.a., of the AD/DA converters, up-/downconverters, bandpass filters). Next, a network of switches connects each RF chain with one of the K ports. Note that each of the K ports represents a different

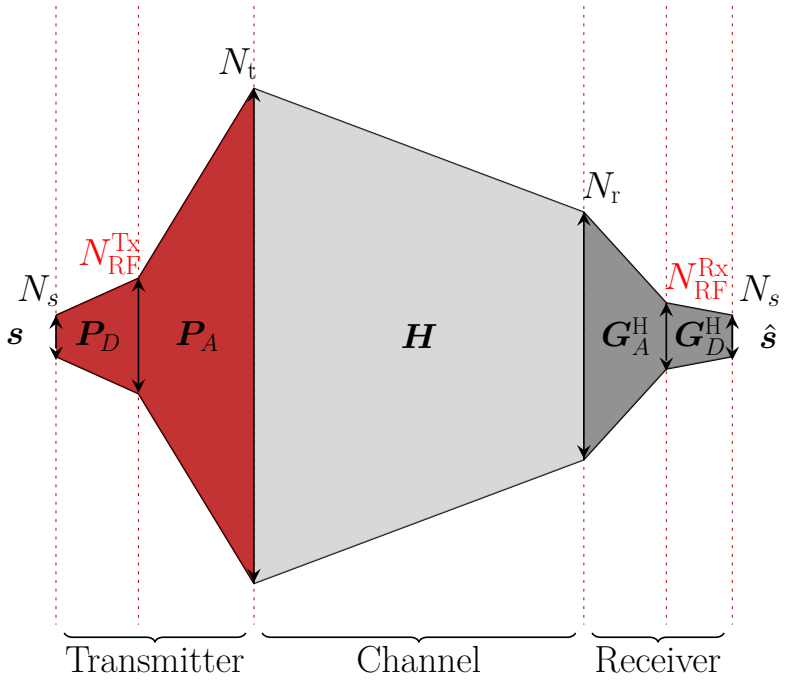

Figure 3.1.. Signal dimensionalities in a system with RF chain reduction

beamforming vector with entries corresponding to the delays between the port and all the antennas.

The lens antenna architecture has a few drawbacks. Firstly, the flexibility of this architecture is limited in two ways—the number of available beamformers is fixed and the freedom of constructing the codebook is limited by physical properties of the design. Secondly, the lens antennas are typically large [20].

An alternative that potentially allows for more flexibility, is the PSN. There, in the Rx, the signal from an antenna undergoes a custom phase shift before being fed into each of the available RF chains (and vice versa in the Tx).

3.4. Partially Interconnected Hybrid Beamforming

The schematics of the fully interconnected phase shifter network (which we refer to in the remainder as to fully interconnected hybrid beamforming (FI-HBF)) is depicted in Fig. 3.2. Such design can significantly reduce the numbers of ADCs but creates a number of new problems.

Table 3.1.. Exemplary power consumption of electronic elements used for the construction of PSN

Device	**Power [mW]**
Power Amplifier	$40 - 250$
Low Noise Amplifier	$4 - 86$
Phase Shifter	$15 - 110$
ADC	$15 - 795$
VCO	$4 - 25$

First, implementing a reliable analog part of FI-HBF on mmWave frequencies is a very big challenge in the design stage, e.g., it requires a lot of effort to keep the PSN synchronized and to avoid crosstalk, especially with the density of wiring required for FI-HBF.

Second, although for the sake of theoretical works the PSN is assumed passive, in fact all the involved elements need power in order to operate. An effort to quantify the overall power budget required for realizing the circuit has been taken in a few works, e.g., in [21,30,57,60]. In Table 3.1 we show an exemplary summary presented in [57] concerning the power consumption of different analog components. We note that the numbers are not very precise and only to a degree compatible with other works, e.g., [21, 60]. Therefore, in this work we do not stick to particular power consumption values and assume that the PSN is passive. We emphasize that this simplification is beneficial for the FI-HBF architecture over SPI-HBF.

Moreover, the electronic components included in the circuit exhibit non-negligible power dissipation. Firstly, all components (splitters, power combiners, phase shifters) have insertion losses determined by their quality. Secondly, the power splitters and combiners can not be realized as (cascades of) 3 port components, if they should be lossless, matched, and reciprocal at the same time—the inherent properties of the S-matrix prohibit 3 port devices from sharing all the properties simultaneously [56]. Consequently, a big portion of power (dependent on the number of inputs) is burnt in the combiner's balast resistors [21, 25].

An answer to the aforementioned problems is a simplification of the FI-HBF structure, namely partially interconnected hybrid beamforming (PI-HBF) where there *exist* RF chains without a connection to some of the antennas. Among the class of PI-HBF architectures, an important subset are the subarray partially interconnected hybrid beamforming (SPI-HBF) architectures, where *disjoint* sets of antennas (sub-

arrays) are connected to *disjoint* sets of RF chains. Moreover, all antennas within a subarray are connected to all the RF chains belonging to a given subarray. Such an architecture is presented in Fig. 3.3 and is the main topic of the publication.

3.5. Subarray Hybrid Beamforming

The analysis of SPI-HBF is important for two reasons. Firstly, it can provide quantitative answers regarding the eventual performance loss if SPI-HBF is implemented instead of FI-HBF. Secondly, it gives an insight into the performance of a transceiver constructed of cooperating FI-HBF blocks.

We note that a special case of FI-HBF is analog beamforming (AB) where only one RF chain is connected (with phase shifters in between) to all the antennas (see Fig. 3.4). Such an architecture has already been successfully implemented in both a testbed setting [15] and as a commercial product [1]. Moreover, it has become a part of the 802.11ad wireless standard.

An important SPI-HBF architecture is the one where the analog stage of the HBF is constructed of multiple analog beamforming blocks, see Fig. 3.5. Moreover, to the best knowledge of the author, it is the only SPI-HBF variant considered in the literature apart from the works of the author of the thesis [35, 36, 38]. The reason for this is twofold. First, this is the simplest SPI-HBF structure, as it does not require (at the Tx) any adders. Second, it simplifies to a large degree the design of the analog precoding.

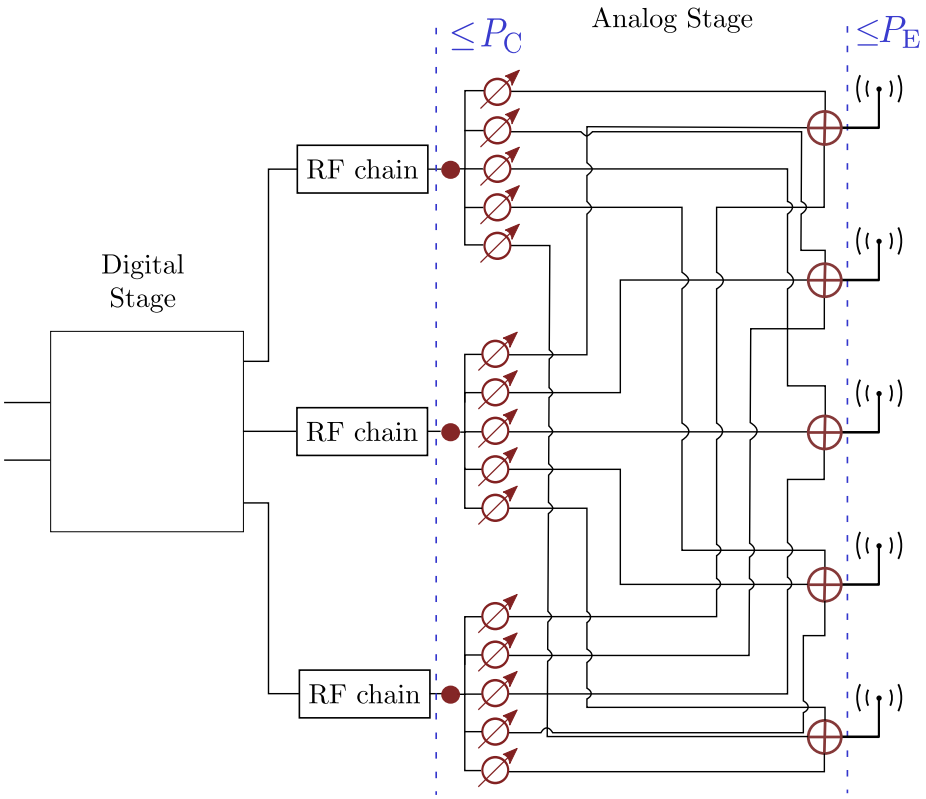

Figure 3.2.. Exemplary fully interconnected (FI-HBF) transmitter structure. $N_{\mathrm{t}} = 5$, $N_{\mathrm{t}}^{\mathrm{RF}} = 3$

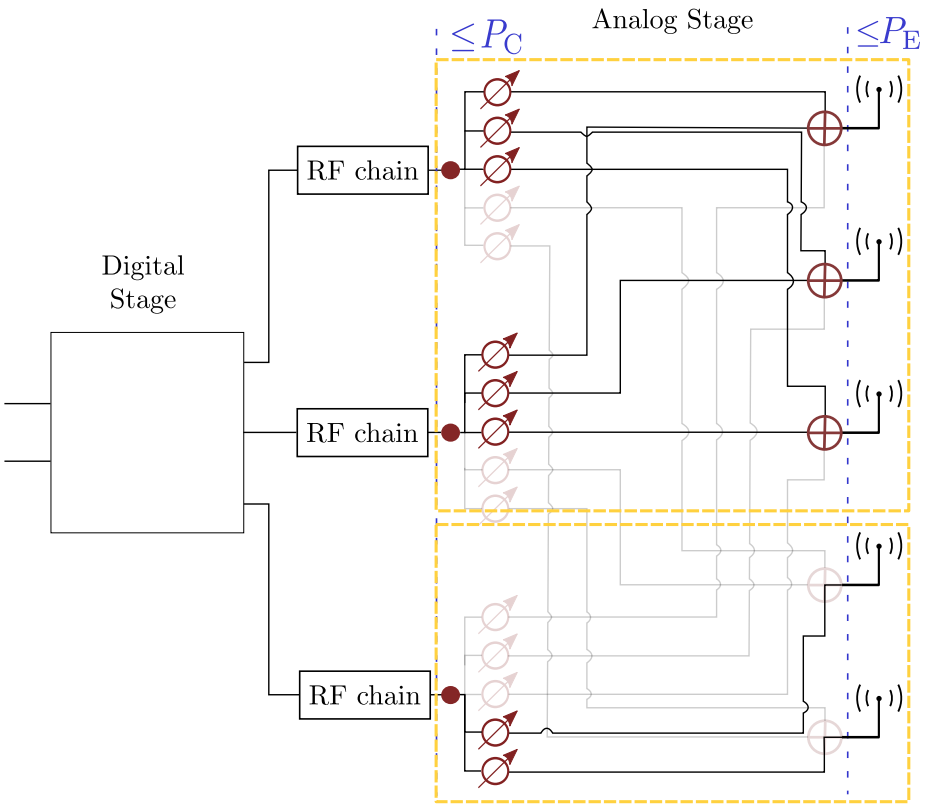

Figure 3.3.. Exemplary subarray partially interconnected (SPI-HBF) transmitter structure. $N_{\mathrm{t}} = 5$, $N_{\mathrm{t}}^{\mathrm{RF}} = 3$, $S_{\mathrm{t}} = 2$, $N_{\mathrm{t}}^{1} = 3$, $N_{\mathrm{t}}^{2} = 2$, $N_{\mathrm{t}}^{\mathrm{RF},1} = 2$, $N_{\mathrm{t}}^{\mathrm{RF},2} = 1$

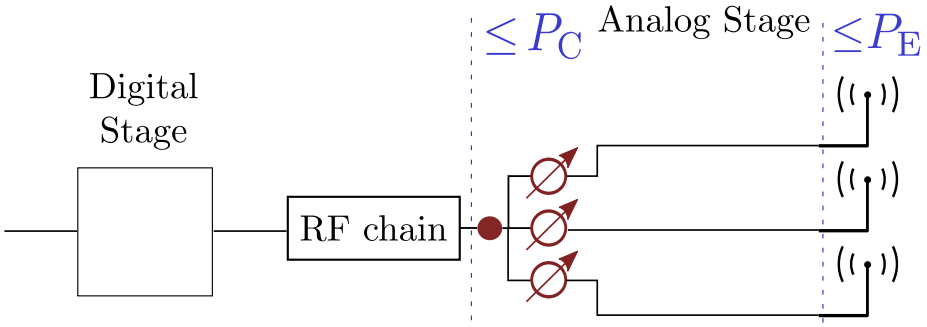

Figure 3.4.. Exemplary Analog Beamforming transmitter structure. $N_{\mathrm{t}} = 3$, $N_{\mathrm{t}}^{\mathrm{RF}} = 1$

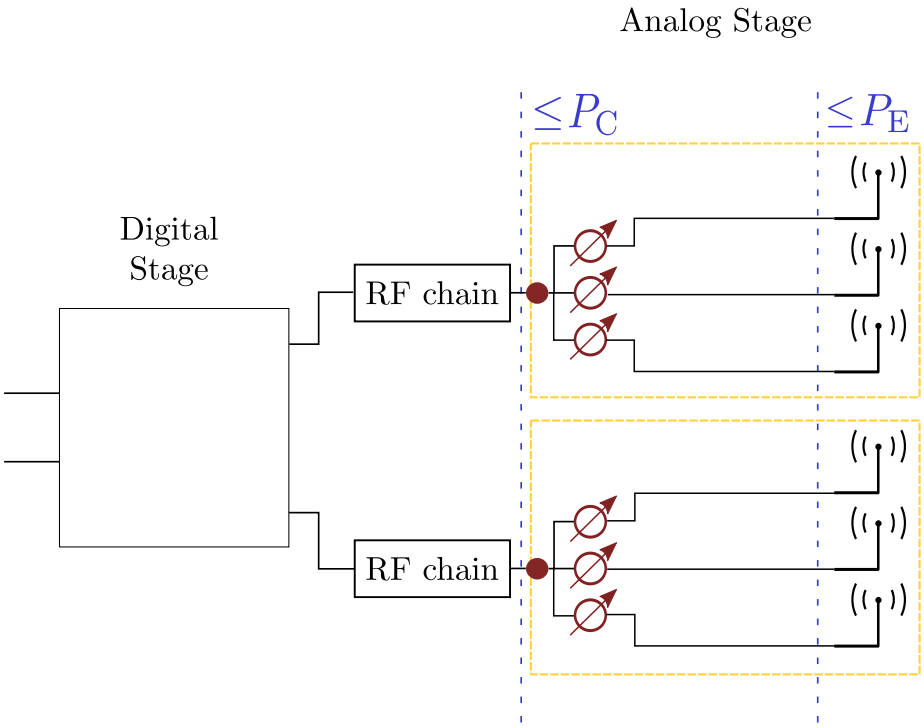

Figure 3.5.. Exemplary SPI-HBF transmitter structure with one RF chain per subarray. $N_{\mathrm{t}} = 6$, $N_{\mathrm{t}}^{\mathrm{RF}} = 2$, $S_{\mathrm{t}} = 2$, $N_{\mathrm{t}}^{1} = 3$, $N_{\mathrm{t}}^{2} = 3$, $N_{\mathrm{t}}^{\mathrm{RF},1} = 1$, $N_{\mathrm{t}}^{\mathrm{RF},2} = 1$

4

System Model

In this chapter, we explain the system model which we consider in the thesis. We formulate the input-output relation of the system and discuss its specifics due to the operating frequency band. Also, we introduce the performance measure used for the system. Furthermore, we discuss how the analog stage of SPI-HBF can be modeled mathematically.

4.1. Input-output relation and ergodic rate

The analog circuit in HBF can be modeled as a linear map between the spaces of dimensions corresponding to the number of RF chains and the number of antennas. The linear maps are described with the matrices $\boldsymbol{G}_A \in \mathbb{C}^{N_\mathrm{r}^\mathrm{RF} \times N_\mathrm{r}}$ and $\boldsymbol{P}_A \in \mathbb{C}^{N_\mathrm{t}^\mathrm{RF} \times N_\mathrm{t}}$ at the Rx and Tx, respectively. Consequently, the system equation for a single link MIMO system with HBF reads

$$\boldsymbol{y} = \boldsymbol{G}_A^\mathrm{H} \left(\boldsymbol{H} \boldsymbol{P}_A \boldsymbol{x} + \boldsymbol{\eta} \right), \tag{4.1}$$

where $\boldsymbol{y} \in \mathbb{C}^{N_\mathrm{r}^\mathrm{RF}}$ is the output vector of the ADCs at the Rx and $\boldsymbol{x} \in \mathbb{C}^{N_\mathrm{t}^\mathrm{RF}}$ is the input vector to the DACs at the Tx, $\boldsymbol{H} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{t}}$ is the channel transfer matrix, and $\boldsymbol{\eta} \sim \mathcal{N}_\mathbb{C}(\boldsymbol{0}, \boldsymbol{R}_\eta)$ represents the Gaussian distributed noise vector with covariance matrix $\boldsymbol{R}_\eta \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{r}}$.

Further, we denote the data signal vector with $\boldsymbol{s} \in \mathbb{C}^{N_\mathrm{s}}$. The data signal is precoded before transmitting with a function $p(\cdot)$, so $\boldsymbol{x} = p(\boldsymbol{s})$ and, respectively, recovered at the receiver with a combining function $c(\cdot)$ such that $\hat{\boldsymbol{s}} = c(\boldsymbol{y})$. In order to reduce

complexity of the transceivers, we resort to linear precoding and combining and write $p(\boldsymbol{s}) = \boldsymbol{P}_D\boldsymbol{s}$ and $\hat{\boldsymbol{s}} = \boldsymbol{G}_D^{\mathrm{H}}\boldsymbol{y}$. From now on, we refer to $\boldsymbol{P}_D$ and $\boldsymbol{G}_D$ as to digital precoding and combining matrices, respectively. Finally, we write the recovered data signal vector as

$$\hat{\boldsymbol{s}} = \boldsymbol{G}_D^{\mathrm{H}}\boldsymbol{y} = \boldsymbol{G}_D^{\mathrm{H}}\boldsymbol{G}_A^{\mathrm{H}}\left(\boldsymbol{H}\boldsymbol{P}_A\boldsymbol{P}_D\boldsymbol{s} + \boldsymbol{\eta}\right). \tag{4.2}$$

The performance measure used throughout the work is the ergodic achievable rate of the system

$$R = \mathrm{E}_{\boldsymbol{H}}[I(\boldsymbol{s}, \hat{\boldsymbol{s}})]. \tag{4.3}$$

We assume i.i.d. proper Gaussian distribution of the data vector $\boldsymbol{s} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{I})$. We take the assumption of block fading and perfect knowledge of the channel matrix $\boldsymbol{H}$ at both the transmitter and receiver in each block. Therefore, for each channel realization we aim at maximizing $R_{\boldsymbol{H}} = I(\boldsymbol{s}, \hat{\boldsymbol{s}}; \boldsymbol{H})$.

Assuming Gaussian signaling, $R_{\boldsymbol{H}}$ takes the following form (c.f. (2.11))

$$R_{\boldsymbol{H}} = \log_2 \det(\boldsymbol{I} + \boldsymbol{R}_{\eta,\mathrm{eff}}^{-1}\boldsymbol{G}_D^{\mathrm{H}}\boldsymbol{H}_{\mathrm{A}}\boldsymbol{P}_D\boldsymbol{P}_D^{\mathrm{H}}\boldsymbol{H}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{G}_D), \tag{4.4}$$

where $\boldsymbol{H}_{\mathrm{A}} = \boldsymbol{G}_A^{\mathrm{H}}\boldsymbol{H}\boldsymbol{P}_A \in \mathbb{C}^{N_{\mathrm{r}}^{\mathrm{RF}} \times N_{\mathrm{t}}^{\mathrm{RF}}}$ is the effective channel gain matrix between the RF chains and $\boldsymbol{R}_{\eta,\mathrm{eff}} = \boldsymbol{G}_D^{\mathrm{H}}\boldsymbol{G}_A^{\mathrm{H}}\boldsymbol{R}_{\eta}\boldsymbol{G}_A\boldsymbol{G}_D$ is the covariance matrix of the noise vector after the combining at the Rx.

4.2. Analog Precoding and Combining Matrices

The operation of PSNs can be modeled by linear operations under the assumption that the combiners and splitters work ideally (i.e., they sum up/divide the signals present at their inputs). Then, the (i, j) entry of $\boldsymbol{P}_A(\boldsymbol{G}_A)$ represents the effect of the path between the ith Tx(Rx) antenna and jth Tx(Rx) RF chain. This includes, i.a., the insertion loss and power consumption of the splitters, phase shifters, and combiners as well as the power dissipation of the combiner (c.f. Sec. 3.4).

In our modeling, we take into account only the latter— this is an unavoidable effect which is not dependent on the particular hardware selected for the implementation.

Consequently, the construction of $\boldsymbol{P}_A$ and $\boldsymbol{G}_A$ for subarrays[1] is as follows

$$|[\boldsymbol{P}_A]_{i,j}| = \begin{cases} \xi_{A,\mathrm{TX}}^{(k)} & \text{iff the } i\text{th antenna and } j\text{th RF chain} \\ & \text{belong to the } k\text{th transmit subarray,} \\ 0 & \text{otherwise,} \end{cases} \tag{4.5a}$$

$$|[\boldsymbol{G}_A]_{i,j}| = \begin{cases} \xi_{A,\mathrm{RX}}^{(l)} & \text{iff the } i\text{th antenna and } j\text{th RF chain} \\ & \text{belong to the } l\text{th receive subarray,} \\ 0 & \text{otherwise.} \end{cases} \tag{4.5b}$$

with

$$\begin{aligned} \xi_{A,\mathrm{TX}}^{(k)} &= \left(N_{\mathrm{t}}^{\mathrm{RF},k} N_{\mathrm{t}}^{k}\right)^{-1/2}, \\ \xi_{A,\mathrm{RX}}^{(l)} &= \left(N_{\mathrm{r}}^{\mathrm{RF},l} N_{\mathrm{r}}^{l}\right)^{-1/2}, \end{aligned} \tag{4.6}$$

where $N_{\mathrm{t}}^{k}(N_{\mathrm{r}}^{l})$ is the number of antennas at the Tx(Rx) subarray with index $k(l)$ and $N_{\mathrm{t}}^{\mathrm{RF},k}(N_{\mathrm{r}}^{\mathrm{RF},l})$ reflects the number of RF chains assigned to the kth Tx (lth Rx) subarray. We note that the formula in (4.6) is the same for both the Tx and Rx. In Tx, the splitter has N_{t}^{k} outputs and therefore the $\sqrt{\frac{1}{N_{\mathrm{t}}^{k}}}$ scaling of the signal. For quantifying the scaling at the combiner, we utilize the model for the Wilkinson power combiner [21,25] and the scaling $\sqrt{\frac{1}{N_{\mathrm{t}}^{\mathrm{RF},k}}}$ follows. In the Rx, the number of combiners and splitters are interchanged, but the overall scaling has the same form.

Such a construction results in a power loss in the analog stage which depends on the coherence of the signals entering the combiner. If the signals are fully coherent (which requires a targeted construction of the precoding (combining) matrices), the entire power entering the combiner is preserved in the output. In this thesis, we consider the general case where no such coherence is enforced. This entails an average loss of power proportional to the number of the inputs to the combiner. Therefore, we differ the average consumed power $P_{\mathrm{C}} = \mathrm{E}[\boldsymbol{P}_D \boldsymbol{s}]$ from the average emitted power $P_{\mathrm{E}} = \mathrm{E}[\boldsymbol{P}_A \boldsymbol{P}_D \boldsymbol{s}]$. They coincide when each subarray is equipped only with one RF chain. Otherwise, $P_{\mathrm{C}} < P_{\mathrm{E}}$ and, moreover, the relation between both has a clean closed form if all the combiners have the same L number of inputs, namely $P_{\mathrm{E}} = P_{\mathrm{C}} \frac{1}{L}$. When all the subarrays are equipped with the same number of RF chains, we have $L = \frac{N_{\mathrm{t}}^{\mathrm{RF}}}{S_{\mathrm{t}}}$ and consequently

$$P_{\mathrm{E}} = P_{\mathrm{C}} \frac{S_{\mathrm{t}}}{N_{\mathrm{t}}^{\mathrm{RF}}}. \tag{4.7}$$

[1] Note that this construction covers also FI-HBF.

With $\mathcal{P}_\mathcal{A}$ and $\mathcal{G}_\mathcal{A}$ we denote feasible sets accounting for the restrictions for the analog precoding and combining matrices as in Eq. (4.5a),(4.5b). Moreover, we assume without loss of generality that the antennas belonging to the same subarray have subsequent indices. Consequently, the analog precoding/combining matrices $\boldsymbol{P}_A$ and $\boldsymbol{G}_A$ have a block-diagonal structure.

From (4.6) it follows that the power dissipation in SPI-HBF and FI-HBF can substantially differ. In most works (e.g., [12,18,53]) this is not taken into account and the emitted power P_E serves for the power constraint. In our work, we advocate to use the consumed power P_C as the power constraint. We argue that then the comparison between FI-HBF and SPI-HBF is more fair because it reflects the complexity vs. flexibility trade-off.

4.3. mmWave Channel Model

A common agreement on the mmWave statistical channel model is still awaited for. While a multitude of measurement campaigns has been conducted, they all suffer from uncertainties that are induced by insufficient channel sounding equipment. Nevertheless, the following assumptions are widely agreed in the community for outdoor urban propagation scenario (e.g., in [5, 12, 62, 63, 72]):

- The number of paths is significantly lower than for the sub-6GHz frequency band.
- The paths propagate in space and time clusters.

Therefore, we choose to consider the geometric channel model, which is compatible with aforementioned assumptions. We use the extended Saleh-Valenzuela model [9, 61] where the MIMO channel sampled at time instance dT_s is written as a superposition of individual, clustered paths as

$$\boldsymbol{H}[d] = \beta\sqrt{\frac{N_\mathrm{r}N_\mathrm{t}}{L}}\sum_{l=1}^{N_\mathrm{cl}}\sum_{r=1}^{N_\mathrm{path}^l}\alpha_{r,l}p(dT_s-\tau_l-\tau_{r,l})\boldsymbol{a}_\mathrm{rx}(\boldsymbol{\theta}_{r,l})\boldsymbol{a}_\mathrm{tx}^\mathrm{H}(\boldsymbol{\phi}_{r,l}), \tag{4.8}$$

where N_cl is the number of time-space clusters, N_path^l is the number of paths in the lth cluster, $\alpha_{r,l}$ is the path gain for the rth path in the lth cluster (including the antenna gain), τ_l is the delay of the lth cluster, $\tau_{r,l}$ is the relative delay w.r.t. the cluster delay of the rth path within the lth cluster. $\boldsymbol{\theta}_{r,l} = [\kappa_{r,l}^\theta, \zeta_{r,l}^\theta]$ is the direction of

arrival (DoA) vector of the rth path within the lth cluster, composed of the elevation angle $\kappa_{r,l}^{\theta}$ and the azimuth angle $\zeta_{r,l}^{\theta}$. $\boldsymbol{\phi}_{r,l} = [\kappa_{r,l}^{\phi}, \zeta_{r,l}^{\phi}]$ is the direction of departure (DoD) vector of the rth path within the lth cluster, composed of the elevation angle $\kappa_{r,l}^{\phi}$ and the azimuth angle $\zeta_{r,l}^{\phi}$. L expresses the path loss between the transmitter and the receiver, and β is a normalization factor such that

$$\mathrm{E}\left[\sum_{d=0}^{D} \|\boldsymbol{H}[d]\|_{\mathrm{F}}^{2}\right] = N_{\mathrm{r}} N_{\mathrm{t}}.$$

The contribution of the rth path in the lth cluster for the channel at the time instance dT_s is evaluated by sampling the transfer function p of the pulse-shaping filter at $dT_s - \tau_l - \tau_r$. The vectors $\boldsymbol{a}_{\mathrm{rx}}$ and $\boldsymbol{a}_{\mathrm{tx}}$ are the antenna array response vectors for the receiver and the transmitter, respectively.

A common choice for the arrangement of antennas are the uniform linear arrays (ULA). The array response vector for ULA reads as

$$\boldsymbol{a}_{\mathrm{ULA}}([\kappa, \zeta]) = \frac{1}{\sqrt{M}}\left[1, e^{j\frac{2\pi}{\lambda}d\sin(\zeta)}, \ldots, e^{j(M-1)\frac{2\pi}{\lambda}d\sin(\zeta)}\right]^{\mathrm{T}} \tag{4.9}$$

where M denotes the number of antennas, λ the wavelength of the transmitted/received wave, and d is the spacing of the antenna elements. In our work, we assume half-wavelength antenna elements spacing $d = \frac{\lambda}{2}$. The response is independent of the elevation angle κ and depends only on the azimuth angle ζ.

For the multicarrier setup, the channel matrix $\boldsymbol{H}_k$ at the kth subcarrier is expressed as

$$\boldsymbol{H}_k = \frac{1}{\sqrt{N_{\mathrm{subc}}}} \sum_{d=0}^{D-1} \boldsymbol{H}[d] \exp\left(\frac{j2\pi k}{N_{\mathrm{subc}}} d\right), \tag{4.10}$$

where we assume that the number of subcarriers is larger than the number of taps in the CIR, i.e., $N_{\mathrm{subc}} > D$.

In the work it is assumed that the mmWave channel (4.8) is known perfectly to the transmitter and the receiver. We remark, however, that the problem of channel estimation itself is a non-trivial one in mmWave systems, due to the large anntenna arrays which are utilized and the hardware constraints. In [7, 22], some solutions exploiting specific mmWave channel characteristics (sparsity [7], low rank [22]) are proposed for narrowband channel estimation.

5

Performance Bounds for SPI-HBF

Before proposing practical precoding schemes for SPI-HBF, we show in this chapter that it is possible to derive a tighter upper bound for the achievable rate of the link than the one usually used in the literature (cf. [12,82]). Deriving a tight upper bound is very important in the process of designing practical algorithms—knowing that the performance is close to limit can focus the efforts on other issues like complexity, flexibility etc.

5.1. Preliminaries

The usual approach to upper bounding the performance of hybrid analog-digital systems is to relax the unit-modulus constraint on the entries of the analog precoding (combining) matrices $\boldsymbol{P}_A$ ($\boldsymbol{G}_A$) (c.f. [12, 19, 82]). This relaxation leaves only the constraint of the limited number of RF chains. Therefore, the rate of such system is upper bounded by the constrained capacity C_{FD} (FD standing for full digital) which is expressed as follows

$$\begin{aligned} R_{\boldsymbol{H}} < C_{\mathrm{FD}} = \max_{\boldsymbol{R}_x} & \log_2 \det\left(\boldsymbol{I} + \boldsymbol{R}_\eta^{-1}\boldsymbol{H}\boldsymbol{R}_x\boldsymbol{H}^{\mathrm{H}}\right) \\ & \text{s.t. } \operatorname{tr}(\boldsymbol{R}_x) \le P_{\mathrm{TX}},\ \operatorname{rank}(\boldsymbol{R}_x) \le N_{\mathrm{t}}^{\mathrm{RF}} \\ & \boldsymbol{R}_x \succeq 0 \end{aligned} \tag{5.1}$$

where P_{TX} is the maximal value of the power emitted on average. In case the average emitted power is constrained, $P_{\mathrm{TX}} = P_{\mathrm{E}}$. Otherwise if the consumed power is

constrained, the emitted power can be in general upper bounded by $P_{\mathrm{TX}} = P_{\mathrm{C}}$, and moreover by $P_{\mathrm{TX}} = P_{\mathrm{C}} \frac{S_{\mathrm{t}}}{N_{\mathrm{t}}^{\mathrm{RF}}}$ if all the subarrays have the same number of RF chains (c.f. (4.7)).

The optimization problem in (5.1) is nonconvex due to the rank constraint, but it yields a closed form solution presented in Appendix A which reads

$$\boldsymbol{R}_{\mathrm{FD}}^{\star} = \boldsymbol{P}_{\mathrm{FD}}^{\star} \boldsymbol{P}_{\mathrm{FD}}^{\star\,\mathrm{H}} \tag{5.2}$$

where $\boldsymbol{P}_{\mathrm{FD}}^{\star} \in \mathbb{C}^{N_{\mathrm{t}} \times N_{\mathrm{t}}^{\mathrm{RF}}} = \boldsymbol{U}_{N_{\mathrm{t}}^{\mathrm{RF}}} \boldsymbol{\Sigma}_{\mathrm{FD}}^{1/2}$ is composed of the matrix $\boldsymbol{U}_{N_{\mathrm{t}}^{\mathrm{RF}}}$ consisting of the $N_{\mathrm{t}}^{\mathrm{RF}}$ eigenvectors of $\boldsymbol{H}^{\mathrm{H}} \boldsymbol{R}_{\eta}^{-1} \boldsymbol{H}$ corresponding to the largest eigenvalues, and $\boldsymbol{\Sigma}_{\mathrm{FD}}$ which is a diagonal matrix consisting of powers allocated to the $N_{\mathrm{t}}^{\mathrm{RF}}$ streams obtained by the waterfilling algorithm [24, 71, 74].

The rank constraint accounts only for the reduced number of RF chains, the remaining constraints are relaxed. Namely, the structure of the analog precoding/combining matrices—the unit-magnitude constraint in the case of FI-HBF and, additionally, the block-diagonal structure of $\boldsymbol{P}_A$ present for SPI-HBF, have not been considered.

We start by introducing notation that helps with explicitly relating the input-output system equation to the diagonal blocks of the $\boldsymbol{P}_A$ and $\boldsymbol{G}_A$ matrices and their digital counterparts. For this sake we write the analog precoding and combining matrices as

$$\boldsymbol{P}_A = \begin{bmatrix} \boldsymbol{P}_A^{(1)} & & \\ & \ddots & \\ & & \boldsymbol{P}_A^{(S_{\mathrm{t}})} \end{bmatrix}, \boldsymbol{G}_A = \begin{bmatrix} \boldsymbol{G}_A^{(1)} & & \\ & \ddots & \\ & & \boldsymbol{G}_A^{(S_{\mathrm{r}})} \end{bmatrix}, \tag{5.3}$$

the digital precoding and combining matrices as

$$\boldsymbol{P}_D = \begin{bmatrix} \boldsymbol{P}_D^{(1)} \\ \vdots \\ \boldsymbol{P}_D^{(S_{\mathrm{t}})} \end{bmatrix}, \quad \boldsymbol{G}_D = \begin{bmatrix} \boldsymbol{G}_D^{(1)} \\ \vdots \\ \boldsymbol{G}_D^{(S_{\mathrm{r}})} \end{bmatrix}, \tag{5.4}$$

and

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_{1,1} & \dots & \boldsymbol{H}_{1,S_{\mathrm{t}}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{H}_{S_{\mathrm{r}},1} & \dots & \boldsymbol{H}_{S_{\mathrm{r}},S_{\mathrm{t}}} \end{bmatrix} = \left[\boldsymbol{H}_{:,1}, \dots, \boldsymbol{H}_{:,S_{\mathrm{t}}} \right] \tag{5.5}$$

where S_r stands for the number of subarrays at the Rx, and $\boldsymbol{H}_{:,k}$ is the channel matrix between the kth Tx subarray and all the Rx antennas. The dimensionality of the blocks in the matrices are determined by the system setup—more specifically, the number of RF chains and antennas assigned to the subarrays:

$$\begin{aligned}
\boldsymbol{P}_A^{(k)} &\in \mathbb{C}^{N_\mathrm{t}^k \times N_\mathrm{t}^{\mathrm{RF},k}}, \\
\boldsymbol{G}_A^{(l)} &\in \mathbb{C}^{N_\mathrm{r}^l \times N_\mathrm{r}^{\mathrm{RF},l}}, \\
\boldsymbol{P}_D^{(k)} &\in \mathbb{C}^{N_\mathrm{t}^{\mathrm{RF},k} \times N_\mathrm{s}}, \\
\boldsymbol{G}_D^{(l)} &\in \mathbb{C}^{N_\mathrm{r}^{\mathrm{RF},l} \times N_\mathrm{s}}, \\
\boldsymbol{H}_{l,k} &\in \mathbb{C}^{N_\mathrm{r}^l \times N_\mathrm{t}^k}, \\
\boldsymbol{H}_{:,k} &\in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{t}^k}.
\end{aligned}$$

We use the introduced notation to rewrite (4.2) as follows

$$\hat{\boldsymbol{s}} = \boldsymbol{G}^\mathrm{H}\boldsymbol{H}\boldsymbol{P}\boldsymbol{s} + \boldsymbol{G}^\mathrm{H}\boldsymbol{\eta} = \boldsymbol{G}^\mathrm{H}\sum_{k=1}^{S_\mathrm{t}} \boldsymbol{H}_{:,k}\boldsymbol{x}_k + \boldsymbol{G}^\mathrm{H}\boldsymbol{\eta} \tag{5.6}$$

where

$$\begin{aligned}
\boldsymbol{G} &= \begin{bmatrix} \boldsymbol{G}^{(1)} \\ \vdots \\ \boldsymbol{G}^{(S_\mathrm{r})} \end{bmatrix} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{s}} \\
\boldsymbol{G}^{(l)} &= \boldsymbol{G}_A^{(l)}\boldsymbol{G}_D^{(l)} \in \mathbb{C}^{N_\mathrm{r}^l \times N_\mathrm{s}}, \\
\boldsymbol{P} &= \begin{bmatrix} \boldsymbol{P}^{(1)} \\ \vdots \\ \boldsymbol{P}^{(S_\mathrm{t})} \end{bmatrix} \in \mathbb{C}^{N_\mathrm{t} \times N_\mathrm{s}}, \\
\boldsymbol{P}^{(k)} &= \boldsymbol{P}_A^{(k)}\boldsymbol{P}_D^{(k)} \in \mathbb{C}^{N_\mathrm{t}^k \times N_\mathrm{s}}, \\
\boldsymbol{x}_k &= \boldsymbol{P}^{(k)}\boldsymbol{s} \in \mathbb{C}^{N_\mathrm{t}^k}.
\end{aligned} \tag{5.7}$$

With the new form of the system equation we gain additional insights into the effects of block-diagonal analog precoding and combining. The expression in (5.6) suggests two alternative perspectives of interpreting the structure, namely as

1) point-to-point MIMO system with special structure of linear precoding and combining matrices,

2) multiple-access channel (MAC) with cooperating transmitters and special structure of precoding and combining matrices.

5.2. Low-SNR Upper Bound

We consider the structure of the precoding matrix $\boldsymbol{P}$. We note that the analysis for the combining matrix $\boldsymbol{G}$ follows the same arguments. From (5.7) we see that it is constructed as a vertical concatenation of S_t matrices $\boldsymbol{P}^{(k)}$. Those submatrices are constructed as a product of the $\boldsymbol{P}_A^{(k)}$ and $\boldsymbol{P}_D^{(k)}$ matrices: the first one describing the operation of the analog stage of the subarray and second one of it's digital counterpart. As a result, the feasible set for the submatrices is constrained by

- the entries of $\boldsymbol{P}_A^{(k)}$ are unit modulus,
- column rank deficiency of $\boldsymbol{P}^{(k)}$ if $N_\mathrm{t}^{\mathrm{RF},k} < \min\{N_\mathrm{t}^k, N_\mathrm{s}\}$, where $N_\mathrm{t}^{\mathrm{RF},k}$ is the number of RF chains at the kth subarray ot the transmitter.
- the power constraint $\|\boldsymbol{P}\|_\mathrm{F}^2 \leq P_\mathrm{TX}$.

To summarize, this means that the effective precoding matrix $\boldsymbol{P}$ is constructed as a concatenation of possibly column rank deficient matrices, each lying in an intricate non-convex set. Moreover, a power constraint limits the Frobenius norm of the matrix.

While the first constraint is common for both FI-HBF and SPI-HBF, the second one is SPI-HBF specific. Therefore, we expect the upper bound C_FD which proved to be tight for FI-HBF to work poorly with SPI-HBF as it ignores the second constraint.

A special case emerges in the low-SNR regime. Below a certain SNR, the optimal number of independent streams $N_\mathrm{s}^\star$ becomes smaller than the number of RF chains at any of the subarrays both at the Tx and Rx:

$$N_\mathrm{s}^\star \leq \min\{N_\mathrm{t}^{\mathrm{RF},1}, \ldots, N_\mathrm{t}^{\mathrm{RF},S_\mathrm{t}}, N_\mathrm{r}^{\mathrm{RF},1}, \ldots, N_\mathrm{r}^{\mathrm{RF},S_\mathrm{r}}\}.$$

In such case, all the submatrices $\boldsymbol{P}^{(k)}$ and $\boldsymbol{G}^{(l)}$ have full column rank and the unit-modulus constraint is the only remaining constraint. Therefore, the achievable rate of the system can be upper-bounded by the point-to-point MIMO capacity with the number of streams limited to $\min\{N_\mathrm{t}^{\mathrm{RF},1}, \ldots, N_\mathrm{r}^{\mathrm{RF},S_\mathrm{r}}\}$:

$$\begin{aligned} R_{\boldsymbol{H},\mathrm{Low}} \leq C_{\boldsymbol{H},\mathrm{Low}} = \max_{\boldsymbol{R}_\mathrm{L}} & \log_2 \det\left(\boldsymbol{I} + \boldsymbol{R}_\eta^{-1}\boldsymbol{H}\boldsymbol{R}_\mathrm{L}\boldsymbol{H}^\mathrm{H}\right) \\ \text{s.t.} \quad & \mathrm{tr}(\boldsymbol{R}_\mathrm{L}) \leq P_\mathrm{TX},\ \boldsymbol{R}_\mathrm{L} \succeq 0, \\ & \mathrm{rank}(\boldsymbol{R}_\mathrm{L}) \leq \min\{N_\mathrm{t}^{\mathrm{RF},1}, \ldots, N_\mathrm{r}^{\mathrm{RF},S_\mathrm{r}}\} \end{aligned} \tag{5.8}$$

where $\boldsymbol{R}_{\mathrm{L}}$ is the transmit covariance matrix optimal for the MIMO channel $\boldsymbol{H}$ given the number of streams limited by $\min\{N_{\mathrm{r}}^{\mathrm{RF},1},\ldots,N_{\mathrm{r}}^{\mathrm{RF},S_{\mathrm{t}}}\}$ and P_{TX} is like in (5.1). The solution to the problem is outlined in Appendix A. Moreover, we write

$$C_{\mathrm{Low}} = \mathrm{E}_{\boldsymbol{H}}\left[C_{\boldsymbol{H},\mathrm{Low}}\right]. \tag{5.9}$$

5.3. High-SNR Upper Bound

The second expression in (5.6) relates the system to a MAC scenario with cooperation and user's input distributions within parametric families

$$\begin{aligned}
&\{p_{\boldsymbol{x}_k}(\boldsymbol{x}_k;\boldsymbol{\theta}) = \mathcal{N}_{\mathbb{C}}\left(\boldsymbol{x}_k;\boldsymbol{0},\boldsymbol{\theta}\right)|\boldsymbol{\theta}\in\Theta_k\}\\
&\Theta_k = \Bigg\{\boldsymbol{P}_A^{(k)}\boldsymbol{P}_D^{(k)}\boldsymbol{P}_D^{(k)\mathrm{H}}\boldsymbol{P}_A^{(k)\mathrm{H}}|\boldsymbol{P}_A^{(k)}\in\mathcal{P}_{\mathcal{A}},\ \boldsymbol{P}_D^{(k)}\in\mathbb{C}^{N_{\mathrm{t}}^{\mathrm{RF},k}\times N_{\mathrm{s}}}\\
&\qquad \|\boldsymbol{\Gamma}\boldsymbol{P}_D\|_{\mathrm{F}}^2 \le P_{\mathrm{TX}}\Bigg\}
\end{aligned}$$

where $\mathcal{N}_{\mathbb{C}}\left(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}\right)$ denotes the multivariate Gaussian density of argument $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Moreover, $(\boldsymbol{\Gamma},P_{\mathrm{TX}}) = (\boldsymbol{I},P_{\mathrm{C}})$ if consumed power is constrained and $(\boldsymbol{\Gamma},P_{\mathrm{TX}}) = (\boldsymbol{P}_A,P_{\mathrm{E}})$ if emitted power is constrained. Moreover, at the receiver the combining matrix $\boldsymbol{G}$ in general does not preserve sufficient statistics about the signal vector $\boldsymbol{s}$.

The cooperative MAC has a similar structure as a regular MAC scenario, but with transmitters sharing the power budget and coordinating to jointly code the information.[1] The rate region for cooperative MAC is a simplex described as follows [80]:

$$\mathcal{R}_{\mathrm{CM}} = \left\{(R_1,\ldots,R_{S_{\mathrm{t}}})\in\mathbb{R}^{S_{\mathrm{t}}}\,\middle|\,\sum_{k=1}^{S_{\mathrm{t}}}R_k \le C_{\boldsymbol{H}},\ R_k\ge 0\ \forall k\right\}$$

with R_k corresponding to the rate of the kth user and

$$C_{\boldsymbol{H}} = \max_{\left\{p_{\boldsymbol{x}_k}(\cdot;\boldsymbol{\theta})\right\}} R_{\boldsymbol{H}} = \max_{\left\{p_{\boldsymbol{x}_k}(\cdot;\boldsymbol{\theta})\right\}} I(\boldsymbol{x}_1\boldsymbol{x}_2,\ldots,\boldsymbol{x}_{S_{\mathrm{t}}},\hat{\boldsymbol{s}})$$

being the capacity of the point-to-point channel $\boldsymbol{H}$ given the constraints imposed on the feasible set of distributions of $\boldsymbol{x}_k$. In general, the value of $C_{\boldsymbol{H}}$ remains unknown.

[1] The analogous cooperative BC is described in [78]

In the following, we show that in high SNR the described MAC setup may be considered as a regular MAC (uncooperative) and based on this observation we show how to tightly upper-bound $C_{\boldsymbol{H}}$.

At high SNR, all the available spatial degrees of freedom are exploited, i.e., DoF = N_s. In a hybrid system,

$$\mathrm{DoF}_\mathrm{HBF} = \mathrm{rank}(\boldsymbol{G}_A^\mathrm{H} \boldsymbol{H} \boldsymbol{P}_A). \tag{5.10}$$

If $\mathrm{DoF}_\mathrm{HBF} = N_\mathrm{t}^\mathrm{RF}$, which is the case if $N_\mathrm{t}^\mathrm{RF} \leq \min\{N_\mathrm{r}^\mathrm{RF}, \mathrm{rank}(\boldsymbol{H})\}$, the optimal digital precoding matrix is square and for SNR $\to \infty$ also unitary $\boldsymbol{P}_D \boldsymbol{P}_D^\mathrm{H} = \boldsymbol{P}_D^\mathrm{H} \boldsymbol{P}_D = \boldsymbol{I}$. Therefore, there exists a finite SNR value for which $\boldsymbol{P}_D^{(k_1)} {\boldsymbol{P}_D^{(k_2)}}^\mathrm{H} = \boldsymbol{0}$ (which is a weaker condition than $\boldsymbol{P}_D$ being unitary). This entails $\mathrm{E}[\boldsymbol{x}_{k_1} \boldsymbol{x}_{k_2}^\mathrm{H}] = \boldsymbol{0},\ \forall k_1 \neq k_2$, i.e., that the data transmitted by different subarrays is uncorrelated.

This allows writing $R_{\boldsymbol{H}}$ as the sum rate of a virtual uncooperative MAC system with user k precoding with $\boldsymbol{P}^{(k)}$ precoding matrix and the receiver using $\boldsymbol{G}$ as the combiner

$$\begin{aligned} R_{\boldsymbol{H},\mathrm{High}} = &- \log_2 \det(\boldsymbol{R}_{\eta,\mathrm{eff}}) + \\ &\log_2 \det \left(\boldsymbol{R}_{\eta,\mathrm{eff}} + \sum_{k=1}^{S_\mathrm{t}} \boldsymbol{G}^\mathrm{H} \boldsymbol{H}_{:,k} \boldsymbol{P}^{(k)} \boldsymbol{P}^{(k)\mathrm{H}} \boldsymbol{H}_{:,k}^\mathrm{H} \boldsymbol{G} \right) \end{aligned} \tag{5.11}$$

where $\boldsymbol{R}_{\eta,\mathrm{eff}}$ defined as in (4.4). We upper bound this expression with the capacity of the MAC channel with limited multiplexing capabilities[2] (as $\boldsymbol{P}^{(k)}$ are in the usual case rank deficient) and a sum power constraint

$$\begin{aligned} R_{\boldsymbol{H},\mathrm{High}} \leq C_\mathrm{High}(\boldsymbol{H}) = &\max_{\boldsymbol{P}^{(1)},\ldots,\boldsymbol{P}^{(S_\mathrm{t})}} \Big\{ - \log_2 \det(\boldsymbol{R}_\eta) + \\ &\log_2 \det \left(\boldsymbol{R}_\eta + \sum_{k=1}^{S_\mathrm{t}} \boldsymbol{H}_{:,k} \boldsymbol{P}^{(k)} \boldsymbol{P}^{(k)\mathrm{H}} \boldsymbol{H}_{:,k}^\mathrm{H} \right) \Big\} \\ \text{s.t.}\ \ &\mathrm{rank}(\boldsymbol{P}^{(k)}) \leq N_\mathrm{t}^{\mathrm{RF},k}\ \forall k \\ &\sum_{k=1}^{S_\mathrm{t}} \mathrm{tr}(\boldsymbol{P}^{(k)} \boldsymbol{P}^{(k)\mathrm{H}}) \leq P_\mathrm{TX} \end{aligned} \tag{5.12}$$

In Algorithm 1, we present the solution from [32], which is suited to determine the capacity expression from (5.12).

[2]Limited multiplexing capabilities refers to a limit imposed on the number of transmitted streams, which is more restrictive than the number of available antennas

Algorithm 1 MAC capacity with limited multiplexing capabilities [32]

Require: $\boldsymbol{H}_{:,k}$, $\boldsymbol{R}_\eta$, P_{TX}

Choose s_0, $s' = 1$
$\boldsymbol{T}^{(k)} \leftarrow \left[\boldsymbol{I}_{N_{\mathrm{t}}^{\mathrm{RF},k}} \; \boldsymbol{0}^{N_{\mathrm{t}}^{\mathrm{RF},k} \times N_{\mathrm{t}}^{k} - N_{\mathrm{t}}^{\mathrm{RF},k}}\right]^{\mathrm{T}}$
$\boldsymbol{Q}^{(k)} \leftarrow \frac{P_{\mathrm{TX}}}{S_{\mathrm{t}}} \boldsymbol{T}^{(k)} \boldsymbol{T}^{(k)\mathrm{H}} \quad \forall k$
$\boldsymbol{\Sigma} \leftarrow \boldsymbol{R}_\eta + \sum_{k=1}^{S_{\mathrm{t}}} \boldsymbol{H}_{:,k} \boldsymbol{Q}^{(k)} \boldsymbol{H}_{:,k}^{\mathrm{H}}$
$C_{\boldsymbol{H},\mathrm{High}}^{\mathrm{last}} \leftarrow \log_2 \det(\boldsymbol{\Sigma})$
repeat
    $\delta \boldsymbol{T}^{(k)} \leftarrow \boldsymbol{H}_{:,k} \boldsymbol{\Sigma}^{-1} \boldsymbol{H}_{:,k}{}^{\mathrm{H}} \boldsymbol{T}^{(k)}$
    $\lambda \leftarrow \sqrt{\frac{P_{\mathrm{TX}}}{\sum_{k=1}^{S_{\mathrm{t}}} \|\delta \boldsymbol{T}^{(k)}\|_{\mathrm{F}}^2}}$
    repeat
        $\boldsymbol{T}^{(k)\prime} \leftarrow \boldsymbol{T}^{(k)} + \frac{s_0}{s'} \lambda \delta \boldsymbol{T}^{(k)} \quad \forall k$
        $\boldsymbol{T}^{(k)\prime} \leftarrow \boldsymbol{T}^{(k)\prime} \sqrt{\frac{P_{\mathrm{TX}}}{\sum_{k=1}^{S_{\mathrm{t}}} \|\boldsymbol{T}^{(k)\prime}\|_{\mathrm{F}}^2}} \quad \forall k$
        $\boldsymbol{\Sigma}' \leftarrow \boldsymbol{R}_\eta + \sum_{k=1}^{S_{\mathrm{t}}} \boldsymbol{H}_{:,k} \boldsymbol{T}^{(k)\prime} \boldsymbol{T}^{(k)\prime\mathrm{H}} \boldsymbol{H}_{:,k}{}^{\mathrm{H}}$
        $C_{\boldsymbol{H},\mathrm{High}}^{\mathrm{new}} \leftarrow \log_2 \det(\boldsymbol{\Sigma}')$
        if $C_{\boldsymbol{H},\mathrm{High}}^{\mathrm{new}} \leq C_{\boldsymbol{H},\mathrm{High}}^{\mathrm{last}}$ **then**
            $s' \leftarrow s' + 1$
        end if
    until $C_{\boldsymbol{H},\mathrm{High}}^{\mathrm{new}} > C_{\boldsymbol{H},\mathrm{High}}^{\mathrm{last}}$
    $\boldsymbol{T}^{(k)} \leftarrow \boldsymbol{T}^{(k)\prime}$, $\boldsymbol{\Sigma} \leftarrow \boldsymbol{\Sigma}'$, $C_{\boldsymbol{H},\mathrm{High}}^{\mathrm{last}} \leftarrow C_{\boldsymbol{H},\mathrm{High}}^{\mathrm{new}}$
until convergence
return $C_{\boldsymbol{H},\mathrm{High}}^{\mathrm{new}}$

We further note that in a point-to-point link, a rate of transmission in one direction can be always achieved with proper precoding also in a reverse link. We exploit that in order to tighten the bound $C_{\mathrm{High}}(\boldsymbol{H})$, which does not account for the subarray structure at the receiver. From

$$R_{\boldsymbol{H},\mathrm{High}} \leq C_{\mathrm{High}}(\boldsymbol{H}), \; R_{\boldsymbol{H}^{\mathrm{H}},\mathrm{High}} \leq C_{\mathrm{High}}(\boldsymbol{H}^{\mathrm{H}})$$

and the aforementioned fact, we get

$$R_{\boldsymbol{H},\mathrm{High}} \leq \min\{C_{\mathrm{High}}(\boldsymbol{H}), C_{\mathrm{High}}(\boldsymbol{H}^{\mathrm{H}})\} := C_{\boldsymbol{H},\mathrm{High}}$$

and we denote the bound averaged over the channel realizations as $C_{\mathrm{High}} = \mathrm{E}_{\boldsymbol{H}} \left[C_{\boldsymbol{H},\mathrm{High}}\right]$.

5.4. Discussion

We note that additional argumentation should be formulated in order to convince the reader that the novel bound from (5.12) is tighter compared to the constrained capacity C_{FD} from (5.1). The usual intuition tells that in high-SNR a MIMO point-to-point link performs equally as a MIMO MAC setup in terms of the (sum) capacity—independent coding of information across the antennas is optimal. Nevertheless, this argumentation is incomplete if the number of degrees-of-freedom is lower than the number of antennas—then the information transmitted by the antennas is not any more independent. This is the case if the number of RF chains is constrained, like in (5.1).

A different situation is highlighted by the arguments presented in this section. Here, the subarray data streams $\boldsymbol{x}_k$ are uncorrelated. This implies that, while similarly as above, the information is not independent throughout all the antennas, there exists some structure—the information transmitted by antennas of different subarrays is uncorrelated. The existence of this particular inter-antenna correlation structure enforces $C_{\mathrm{FD}} \geq C_{\mathrm{High}}$.

5.5. Numerical results

We consider a scenario with a non-symmetric Tx-Rx pair. Namely, the Tx is a base station with significantly more antennas than the Rx, which is considered to be a device with lower complexity. In all simulations, we consider ideal knowledge about the CSI at the Tx. Therefore, the rate is identical in both forward and reverse links and we do not duplicate the simulations.

The Tx and Rx are equipped with 128 and 32 antennas, respectively. All devices are equipped with $N_{\mathrm{RF}} = 8$ RF chains. Without loss of generality, we assume always that all subarrays are identically configured. I.e., each has the same number of RF chains $N_{\mathrm{t}}^{\mathrm{RF},k}(N_{\mathrm{r}}^{\mathrm{RF},l})$ and antennas $N_{\mathrm{t}}^{k}(N_{\mathrm{r}}^{l})$.

In the simulations, we assume a geometric narrowband channel model

$$\boldsymbol{H} = \sqrt{\frac{N_{\mathrm{r}}N_{\mathrm{t}}}{\sum_{c=1}^{N_{\mathrm{cl}}} N_{\mathrm{path}}^{c}}} \sum_{c=1}^{N_{\mathrm{cl}}} \sum_{r=1}^{N_{\mathrm{path}}^{c}} \alpha_{r,c} \boldsymbol{a}_{\mathrm{RX}}(\theta_{r,c}^{\mathrm{RX}}) \boldsymbol{a}_{\mathrm{TX}}^{\mathrm{H}}(\theta_{r,c}^{\mathrm{TX}}) \tag{5.13}$$

where the subarrays together form a uniform linear array (ULA) with half-wavelength antenna spacing. In the simulations, we assume $N_{\mathrm{cl}} = 8$ clusters each containing

$N_{\text{path}}^{l} = 10$ paths. The angles of arrival/departure are drawn from Laplacian distribution with the central angle uniformly distributed over the $[0,\ 2\pi]$ interval and spread of 7.5°. We generate 1000 independent channel realizations.

In our plots, we assume that the noise variance at the Rx is uncorrelated across the antennas with $\boldsymbol{R}_{\eta} = \boldsymbol{I}$ [mW] and that there is no path loss. We note, that in such case the emitted power coincides with the Rx SNR.

We show how the novel upper bounds (for the low- and high-SNR region) compare with the constrained capacity (5.1). We examine the results in Figs. 5.1-5.4 for consumed power constraint and 5.5 for emitted power constraint. We conclude, that with increasing number of subarrays, the bounds diverge from the constrained capacity. This is a desired trend and the results provide an insight into the inevitable performance loss due to the hardware simplification. We follow with an observation that the slope of the high-SNR upper bounds curves (corresponding to the degrees-of-freedom) changes only slightly. This is attributed to the sparse nature of the mmWave channel. However, for more comprehensive evaluation of the usefulness of the upper bounds, a comparison with the achievable rates is required. We cover this in the next chapter.

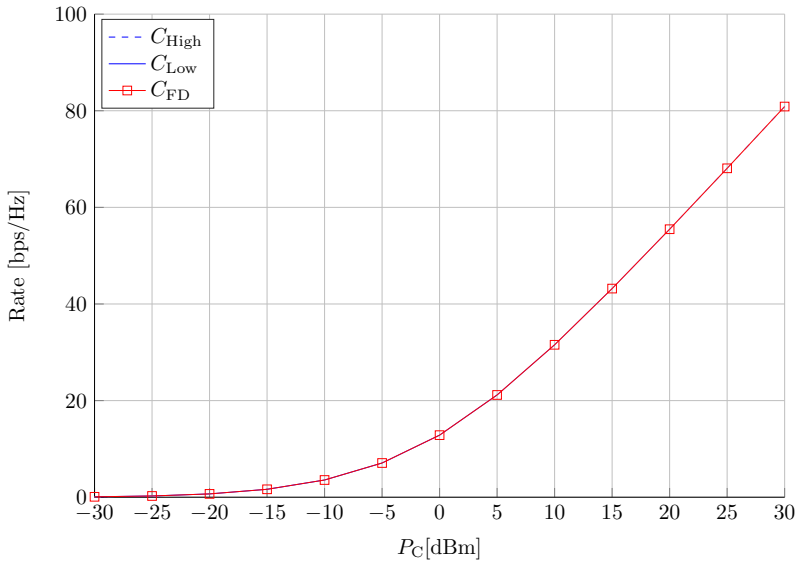

Figure 5.1.. Comparison between the constrained capacity and the novel upper bounds: high SNR upper bound C_{High} and low SNR upper bound C_{Low}. The setup considers $S_{\text{t}} = 1, S_{\text{r}} = 1$ subarray configuration, i.e., a FI-HBF setup, with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx. The consumed power is constrained. In this case, all three curves coincide.

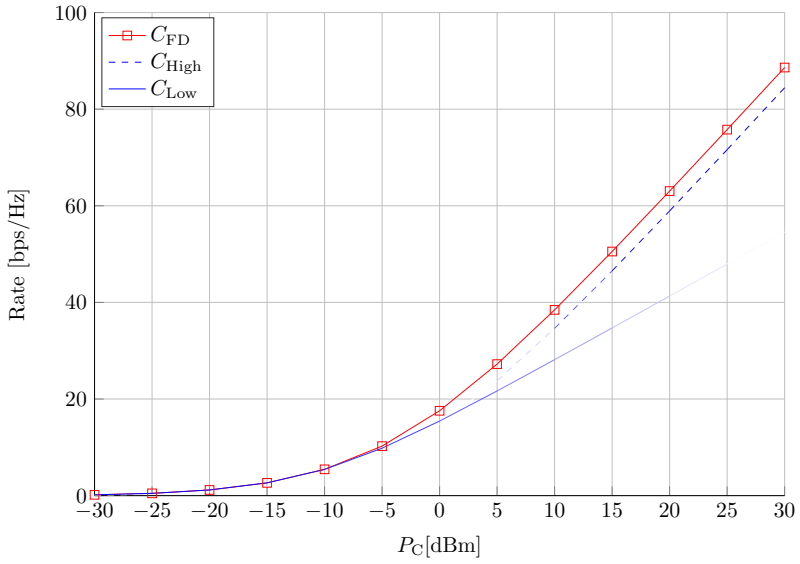

Figure 5.2.. Comparison between the constrained capacity and the novel upper bounds: high SNR upper bound C_{High} and low SNR upper bound C_{Low}. The setup considers $S_{\text{t}} = 2, S_{\text{r}} = 2$ subarray configuration with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx.

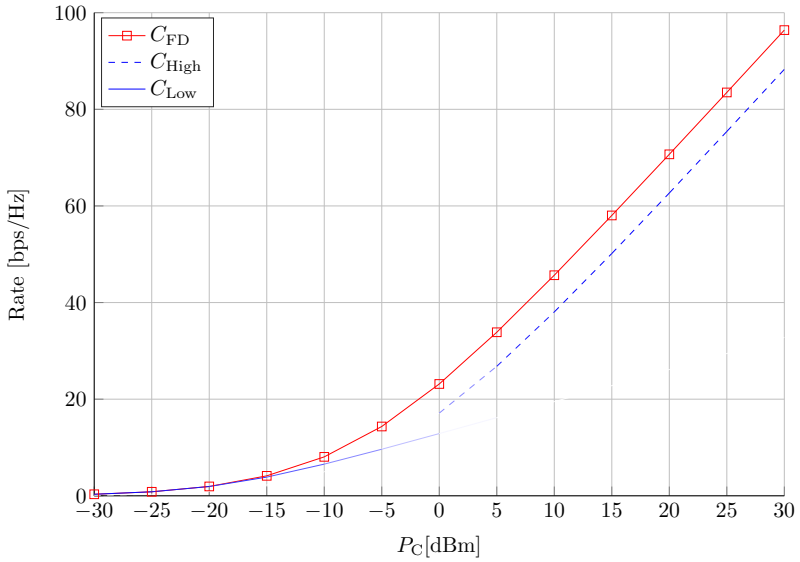

Figure 5.3.. Comparison between the constrained capacity and the novel upper bounds: high SNR upper bound C_{High} and low SNR upper bound C_{Low}. The setup considers $S_{\text{t}} = 4, S_{\text{r}} = 4$ subarray configuration with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx.

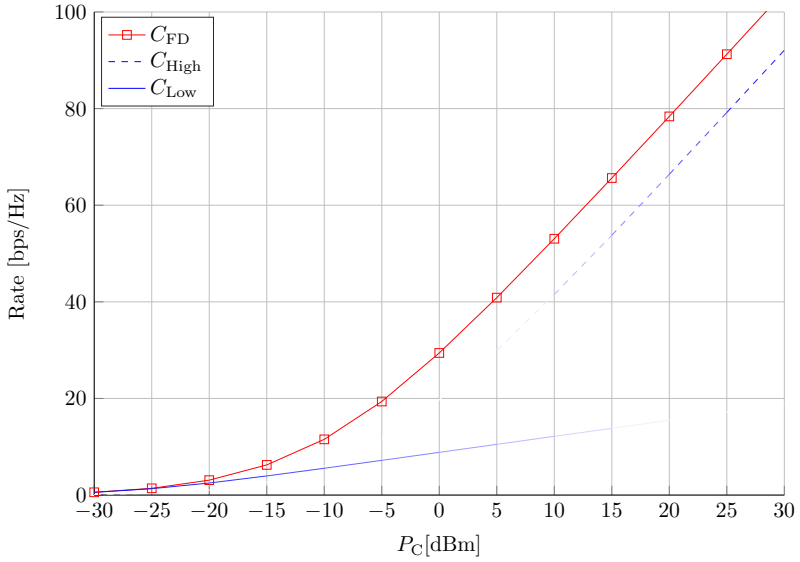

Figure 5.4.. Comparison between the constrained capacity and the novel upper bounds: high SNR upper bound C_{High} and low SNR upper bound C_{Low}. The plot considers $S_{\text{t}} = 8, S_{\text{r}} = 8$ subarray configuration with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx.

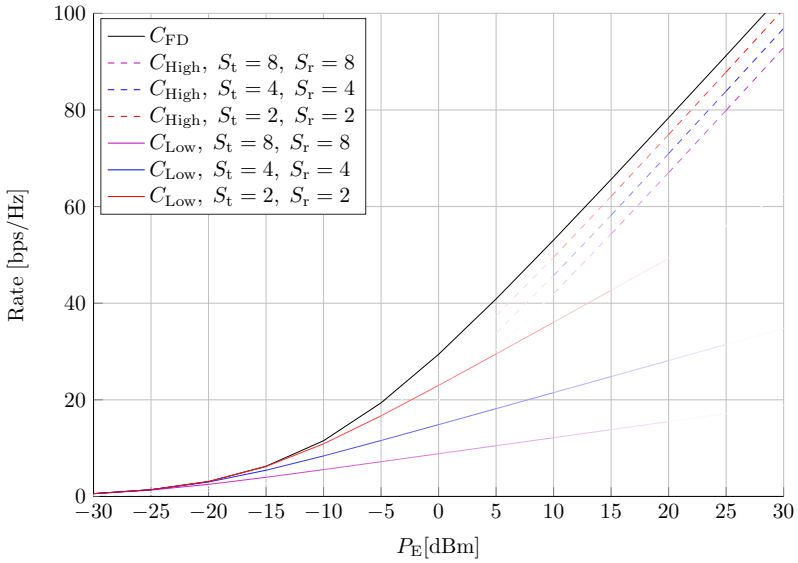

Figure 5.5.. Comparison between the constrained capacity and the novel upper bounds: high SNR upper bound and low SNR upper bound. The curves consider different Tx subarray configurations. The setup considers 128 Tx antennas, 32 Rx antennas, 8 RF chains at the Tx and Rx.

6

Precoding for SPI-HBF

In this chapter, we present the most common approaches to precoding for HBF architectures. We show novel, optimal ways of optimizing the digital precoder in schemes that rely on alternate optimization, minimizing the difference of the hybrid precoding and optimal unconstrained precoding. Furthermore, we propose a greedy precoding scheme which is specially suited for SPI-HBF. Also, we propose a precoding algorithm which is specially designed for scenarios with low-SNR.

In general, the derivation of precoding schemes for SPI-HBF can be divided into two groups. First, there exists significant work on precoding for FI-HBF which can be directly applied to SPI-HBF.

Second, we showed in the previous chapter how the SPI-HBF relates to a (virtual) multiuser system. Such perspective can be utilized in the precoder design. We present solutions that follow this path and provide not only good performance, but also desirable properties from the system perspective.

6.1. State-of-the-Art Overview

The literature on precoding for hybrid beamforming can be separated into a number of overlapping categories, some of which we highlight in the following.

First, in terms of the considered frequency band—hybrid beamforming (under the term of soft antenna selection) has been introduced in [86] for lower frequencies and a few publications followed this idea [70, 77]. Hybrid beamforming for mmWave frequencies is much wider considered, e.g., in [2, 4, 11, 12, 18, 28, 29, 31, 33, 34, 43, 46,

50, 51, 54, 55, 65, 66, 68, 69, 81, 82, 89].

The solutions proposed for mmWaves differ in terms of the degree of being implementation-ready, ranging from practical (and not complex from the theoretical point of view) solutions proposed for the WirelessHD and IEEE 802.11ad [2] standards, through theoretical works which emphasize the practicality of the proposed solutions (c.f., [4, 19, 65, 70]) to works with often idealistic assumptions which provide information about the performance limits of the technology (c.f., [12, 54, 67, 68, 76, 82]).

The next differentiation comes from the specific hybrid architecture considered, starting from the simplest analog beamforming structure (c.f., [2, 77]), through sub-array hybrid beamforming structures with one RF chain per subarray (c.f., [11, 19, 28, 29, 53, 82]) and subarray structures allowing for subarrays with multiple RF chains (c.f. [35, 36, 38]) to solutions for fully interconnected structures (c.f. [4, 12, 31, 76]).

Moreover, there are three main paths of approaching the precoding problem. The first is the codebook beam training approach. There, the goal is to design a protocol which determines the best transmitter-receiver beam pair (where the beams are predefined and form a beam codebook). This approach is adopted in the IEEE 802.11ad standard [2] and covered in a number of publications, e.g., [7, 73, 88]. An approach which seeks for a more optimal solution attempts to find the most accurate approximation of the unconstrained (digital) precoder. This approach forms basis for e.g., [12, 22, 34, 53, 82]. Moreover, some works attempt direct maximization of the rate expression (c.f. [35, 67, 68]).

Finally, there is a number of system assumptions that differentiate the works. This includes the wideband/narrowband assumption, single-/multiuser systems, constraints on the operation of phase shifters (quantized/continuous phase), the degree of making realistic assumptions while modeling the analog elements, etc. The large matrix of assumptions which differentiate the works is expected as hybrid precoding in its core has been designed as a practical solution to a problem. Therefore, introducing additional practical issues is natural. Unfortunately, this also makes comparison between different ideas difficult and often impossible. In the following, we identify the most successful approaches, expand on them and use them as a benchmark to verify the proposed solutions.

6.2. Decomposition of the Unconstrained Precoder

We note that in this section, we concentrate solely on the precoding part. For combining, the problem can be formulated in the same way, except for the power constraint. Any scaling of the combining matrix cannot change the achievable rate of the system, as both the signal and noise part are equally scaled.

A common heuristic approach for maximizing the rate of a hybrid system involves the decomposition of the unconstrained optimal precoder $\boldsymbol{P}_{\mathrm{FD}}^{\star}$ into the analog and digital part. It has been proposed in [12] to use the Frobenius norm of the approximation error as the measure of the decomposition accuracy

$$d_0(\boldsymbol{P}_A, \boldsymbol{P}_D; \boldsymbol{P}_{\mathrm{FD}}^{\star}) = \|\boldsymbol{P}_{\mathrm{FD}}^{\star} - \boldsymbol{P}_A \boldsymbol{P}_D\|_{\mathrm{F}}^2 \tag{6.1}$$

and to minimize the metric by choosing $\boldsymbol{P}_A$ and $\boldsymbol{P}_D$ accordingly

$$\begin{aligned} \{\boldsymbol{P}_A^{\star,0}, \boldsymbol{P}_D^{\star,0}\} = \underset{\boldsymbol{P}_A, \boldsymbol{P}_D}{\arg\min}\ & d_0(\boldsymbol{P}_A, \boldsymbol{P}_D; \boldsymbol{P}_{\mathrm{FD}}^{\star}), \\ \text{s.t.}\ & \|\boldsymbol{\Gamma} \boldsymbol{P}_D\|_{\mathrm{F}}^2 \leq P_{\mathrm{TX}} \\ & \boldsymbol{P}_A \in \mathcal{P}_{\mathcal{A}}. \end{aligned} \tag{6.2}$$

We note that we defined earlier $(\boldsymbol{\Gamma}, P_{\mathrm{TX}}) = (\boldsymbol{I}, P_{\mathrm{C}})$ if the consumed power is constrained and $(\boldsymbol{\Gamma}, P_{\mathrm{TX}}) = (\boldsymbol{P}_A, P_{\mathrm{E}})$ if the emitted power is constrained.

Such approach is justified by arguing that $d_0(\boldsymbol{P}_A, \boldsymbol{P}_D; \boldsymbol{P}_{\mathrm{FD}}^{\star})$ represents the chordal distance between the subspaces represented by the columnspaces of $\boldsymbol{P}_{\mathrm{FD}}^{\star}$ and $\boldsymbol{P}_A \boldsymbol{P}_D$, given the matrix columns are orthogonal. The authors in [12] observe that this is approximately met if the number of antennas at the transmitter is large.

An interesting extension to this approach has been presented in [34]. It has been noticed that any precoding matrix of the form $\boldsymbol{P}_{\mathrm{FD}}^{\star} \boldsymbol{\Psi}$ is also optimal (both in the sense of maximizing the rate and minimizing the MSE) given $\boldsymbol{\Psi}$ is an unitary matrix. Consequently, a more general form of the measure d_0 has been proposed

$$d_1(\boldsymbol{P}_A, \boldsymbol{P}_D, \boldsymbol{\Psi}; \boldsymbol{P}_{\mathrm{FD}}^{\star}) = \|\boldsymbol{P}_{\mathrm{FD}}^{\star} \boldsymbol{\Psi} - \boldsymbol{P}_A \boldsymbol{P}_D\|_{\mathrm{F}}^2 \tag{6.3}$$

and an extension to the optimization problem with an additional optimization vari-

able followed

$$\begin{aligned}\{\boldsymbol{P}_A^{\star,1}, \boldsymbol{P}_D^{\star,1}, \boldsymbol{\Psi}^\star\} = &\underset{\boldsymbol{P}_A, \boldsymbol{P}_D, \boldsymbol{\Psi}}{\arg\min}\ d_1(\boldsymbol{P}_A, \boldsymbol{P}_D, \boldsymbol{\Psi}; \boldsymbol{P}_{\mathrm{FD}}^\star), \\ &\text{s.t. } \boldsymbol{P}_A \in \mathcal{P}_{\mathcal{A}},\ \boldsymbol{\Psi} \in \mathcal{U}_{N_\mathrm{s}} \\ &\|\boldsymbol{\Gamma}\boldsymbol{P}_D\|_\mathrm{F}^2 \leq P_{\mathrm{TX}} \end{aligned} \tag{6.4}$$

with $\mathcal{U}_{N_\mathrm{s}}$ representing the set of all $N_\mathrm{s} \times N_\mathrm{s}$ unitary matrices.

In other words, minimizing d_1 results not only in finding $\boldsymbol{P}_A$ and $\boldsymbol{P}_D$ which approximate most accurately one of the optimal unconstrained precoding matrices. Moreover, it searches also for *an* optimal unconstrained precoding matrix which can be *best approximated* given the constraints. Consequently, we note that $d_1(\boldsymbol{P}_A^{\star,1}, \boldsymbol{P}_D^{\star,1}, \boldsymbol{\Psi}^\star; \boldsymbol{P}_{\mathrm{FD}}^\star) \leq d_0(\boldsymbol{P}_A^{\star,0}, \boldsymbol{P}_D^{\star,0}; \boldsymbol{P}_{\mathrm{FD}}^\star)$, with equality if $\boldsymbol{\Psi}^\star = \boldsymbol{I}$.

In the following, we concentrate on the optimization problem (6.2) and note that the extension to (6.4) follows with the method described in [34].

The optimization problem (6.2) is usually tackled via alternating optimization methods, i.e., updating one of the optimization variables while keeping the other one fixed. In the following, we sketch some of the approaches reported in the literature.

A codebook-based solution has been proposed in [12]. In each alternating step, a subsequent column of the analog precoding matrix $\boldsymbol{P}_A$ is selected from a predefined codebook using the orthogonal matching pursuit (OMP) algorithm. The digital precoding matrix $\boldsymbol{P}_D$ is updated such that the unconstrained objective is minimized $\boldsymbol{P}_D = \boldsymbol{P}_A^\dagger \boldsymbol{P}_{\mathrm{FD}}^\star$. In the end, $\boldsymbol{P}_D$ is scaled with α, where $\alpha = \frac{\sqrt{P_\mathrm{E}}}{\|\boldsymbol{P}_D \boldsymbol{P}_A\|_\mathrm{F}}$ is a scaling factor ensuring that the emitted power constraint is fulfilled. We summarize the algorithm in Algorithm 2.

In [22], an approach inspired by the block coordinate descent (BCD) algorithm has been proposed. In each alternating step the analog precoding matrix $\boldsymbol{P}_A$ is updated via Euclidian projection of the optimal solution $\boldsymbol{P}_D^\dagger \boldsymbol{P}_{\mathrm{FD}}^\star$ onto the feasible set $\mathcal{P}_{\mathcal{A}}$. The digital precoding matrix $\boldsymbol{P}_D$ is updated like in [12]. We summarize the algorithm in Algorithm 3.

Another alternating approach has been reported in [82]. There, the authors notice that the set of feasible analog precoding matrices ($\mathcal{P}_{\mathcal{A}}$) defines a Riemannian manifold. In each alternating step they move along the steepest gradient on the tangent space of the manifold and update the analog precoding matrix $\boldsymbol{P}_A$ by retracting back onto the manifold. The digital precoding matrix $\boldsymbol{P}_D$ is updated like in [12]. Compared to the previous approaches, this one is the most computationally expensive,

Algorithm 2 The OMP precoding algorithm

Require: $\boldsymbol{P}_{\mathrm{FD}}^{\star}$, $\boldsymbol{D}$ (Dictionary)

$\boldsymbol{P}_{\mathrm{res}} = \boldsymbol{P}_{\mathrm{FD}}^{\star}$, $\boldsymbol{P}_A = \mathrm{Null}$
for $i \leq N_{\mathrm{t}}^{\mathrm{RF}}$ **do**
  $\boldsymbol{\Psi} = \boldsymbol{D}^{\mathrm{H}} \boldsymbol{P}_{\mathrm{res}}$
  $k = \arg\max_j \left[\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{H}}\right]_{(j,j)}$
  $\boldsymbol{P}_A = \left[\boldsymbol{P}_A | \boldsymbol{D}_{(:,k)}\right]$
  $\boldsymbol{P}_D = \boldsymbol{P}_A^{\dagger} \boldsymbol{P}_{\mathrm{FD}}^{\star}$
  $\boldsymbol{P}_{\mathrm{res}} = \frac{\boldsymbol{P}_{\mathrm{FD}}^{\star} - \boldsymbol{P}_A \boldsymbol{P}_D}{\|\boldsymbol{P}_{\mathrm{FD}}^{\star} - \boldsymbol{P}_A \boldsymbol{P}_D\|_{\mathrm{F}}}$
end for
$\boldsymbol{P}_D = \frac{\sqrt{P_{\mathrm{E}}}}{\|\boldsymbol{P}_D \boldsymbol{P}_A\|_{\mathrm{F}}} \boldsymbol{P}_D$
return $\boldsymbol{P}_A, \boldsymbol{P}_D$

Algorithm 3 The BCD algorithm

Require: $\boldsymbol{P}_{\mathrm{FD}}^{\star}$

Randomly initialize $\boldsymbol{P}_A$
while not convergence **do**
  $\boldsymbol{P}_D = \boldsymbol{P}_A^{\dagger} \boldsymbol{P}_{\mathrm{FD}}^{\star}$
  $\boldsymbol{P}_A = \Pi_A \left(\boldsymbol{P}_{\mathrm{FD}}^{\star} \boldsymbol{P}_D^{\dagger}\right)$
  where $\Pi_A(\cdot)$ is an orthogonal projection into $\mathcal{P}_{\mathcal{A}}$
end while
$\boldsymbol{P}_D = \frac{\sqrt{P_{\mathrm{E}}}}{\|\boldsymbol{P}_D \boldsymbol{P}_A\|_{\mathrm{F}}} \boldsymbol{P}_D$
return $\boldsymbol{P}_A, \boldsymbol{P}_D$

but provides the best results.

It is also worth noticing that it has been shown in [85] and [67] how to construct $\boldsymbol{P}_A$ and $\boldsymbol{P}_D$ such that $d_0 = 0$, given $N_{\mathrm{t}}^{\mathrm{RF}} \geq 2N_{\mathrm{s}}$. In other words, hybrid precoding can perform identically as the unconstrained precoding given sufficient numbers of RF chains are available.

6.2.1. Digital Precoding Matrix Update

In the above works, the update of $\boldsymbol{P}_A$ in the alternating steps has been dominating the discussion. The update of $\boldsymbol{P}_D$ has been either not elaborated on (cf. [12]) or bounds quantifying the non-optimal update of $\boldsymbol{P}_D$ have been derived (cf. [82]). In

the following, we show how to optimally update $\boldsymbol{P}_D$, so that the power constraint is satisfied.

First, we observe that the optimization problem

$$\min_{\boldsymbol{P}_D} \|\boldsymbol{P}_{\mathrm{FD}}^{\star} - \boldsymbol{P}_A \boldsymbol{P}_D\|_{\mathrm{F}}^2, \quad \text{s.t.} \ \ \|\boldsymbol{\Gamma} \boldsymbol{P}_D\|_{\mathrm{F}}^2 \leq P_{\mathrm{TX}} \tag{6.5}$$

is a least-squares problem with a quadratic constraint. For sake of presentation, we rewrite it to a common form

$$\min_{\boldsymbol{p}_{\mathrm{D}}} \|\boldsymbol{p}_{\mathrm{FD}}^{\star} - \boldsymbol{K}_{\mathrm{A}} \boldsymbol{p}_{\mathrm{D}}\|_2^2, \quad \text{s.t.} \ \ \|\boldsymbol{\Gamma}_1 \boldsymbol{p}_{\mathrm{D}}\|_2^2 \leq P_{\mathrm{TX}} \tag{6.6}$$

where $\boldsymbol{p}_{\mathrm{FD}}^{\star} = \mathrm{vec}(\boldsymbol{P}_{\mathrm{FD}}^{\star})$, $\boldsymbol{p}_{\mathrm{D}} = \mathrm{vec}(\boldsymbol{P}_D)$, and $\boldsymbol{K}_{\mathrm{A}} = \boldsymbol{I}_{N_{\mathrm{s}}} \otimes \boldsymbol{P}_A$ where $\otimes$ stands for the Kronecker product. If the emitted power is constrained, $(\boldsymbol{\Gamma}_1, P_{\mathrm{TX}}) = (\boldsymbol{K}_{\mathrm{A}}, P_{\mathrm{E}})$ and $(\boldsymbol{\Gamma}_1, P_{\mathrm{TX}}) = (\boldsymbol{I}, P_{\mathrm{C}})$ for the constraint on the consumed power. The optimal value of the optimization variable is denoted with $\boldsymbol{P}_D^{\star}$ and its vectorized version as $\boldsymbol{p}_{\mathrm{D}}^{\star} = \mathrm{vec}(\boldsymbol{P}_D^{\star})$. The optimization problem is convex, therefore Karush-Kuhn-Tucker (KKT) conditions are sufficient to find the global optimum. After reformulations, they read

$$\begin{aligned}
&\boldsymbol{p}_{\mathrm{D}}^{\star} = \left(\boldsymbol{K}_{\mathrm{A}}^{\mathrm{H}} \boldsymbol{K}_{\mathrm{A}} + \lambda \boldsymbol{\Gamma}_1^{\mathrm{H}} \boldsymbol{\Gamma}_1\right)^{-1} \boldsymbol{K}_{\mathrm{A}}^{\mathrm{H}} \boldsymbol{p}_{\mathrm{FD}}^{\star} && \text{Stationarity} \\
&\|\boldsymbol{\Gamma}_1 \boldsymbol{p}_{\mathrm{D}}^{\star}\|_2^2 \leq P_{\mathrm{TX}} && \text{Primal feasibility} \\
&\lambda \geq 0 && \text{Dual feasibility} \\
&\lambda(\|\boldsymbol{\Gamma}_1 \boldsymbol{p}_{\mathrm{D}}^{\star}\|_2^2 - P_{\mathrm{TX}}) = 0 && \text{Complementary} \\
& && \text{Slackness}
\end{aligned} \tag{6.7}$$

This set of equations can be solved in two steps. First, we check whether the unconstrained solution (corresponding to $\lambda = 0$) is within the feasible set, i.e., if $\|\boldsymbol{\Gamma}_1 \boldsymbol{K}_{\mathrm{A}}^{\dagger} \boldsymbol{p}_{\mathrm{FD}}^{\star}\|_2^2 \leq P_{\mathrm{TX}}$. If so, $\boldsymbol{p}_{\mathrm{D}}^{\star} = \boldsymbol{K}_{\mathrm{A}}^{\dagger} \boldsymbol{p}_{\mathrm{FD}}^{\star}$ (translating to $\boldsymbol{P}_D^{\star} = \boldsymbol{P}_A^{\dagger} \boldsymbol{P}_{\mathrm{FD}}^{\star}$) is the optimal solution. If this is not the case, the constraint is active, i.e., $\|\boldsymbol{\Gamma}_1 \boldsymbol{p}_{\mathrm{D}}^{\star}\|_2^2 = P_{\mathrm{TX}}$.

We continue by treating separately the two power constraints. For the emitted power constraint we have $P_{\mathrm{TX}} = P_{\mathrm{E}}$ and $\boldsymbol{\Gamma}_1 = \boldsymbol{K}_{\mathrm{A}}$. From the stationarity constraint, we have therefore

$$\boldsymbol{p}_{\mathrm{D}}^{\star} = \frac{1}{1+\lambda} \boldsymbol{K}_{\mathrm{A}}^{\dagger} \boldsymbol{p}_{\mathrm{FD}}^{\star}.$$

This solution is therefore a scaled version of the unconstrained solution $\boldsymbol{K}_{\mathrm{A}}^{\dagger} \boldsymbol{p}_{\mathrm{FD}}^{\star}$. The scaling must be such that the constraint is met with equality, i.e.,

$$\|\boldsymbol{K}_{\mathrm{A}} \boldsymbol{p}_{\mathrm{D}}^{\star}\|_2^2 = P_{\mathrm{E}}.$$

From this, we derive the scaling of the unconstrained precoder $\frac{1}{1+\lambda}$ as follows

$$\frac{1}{1+\lambda}\|\boldsymbol{K}_{\mathrm{A}}\boldsymbol{K}_{\mathrm{A}}^{\dagger}\boldsymbol{p}_{\mathrm{FD}}^{\star}\|_2 = \sqrt{P_{\mathrm{E}}}$$
$$\frac{1}{1+\lambda} = \frac{\sqrt{P_{\mathrm{E}}}}{\|\boldsymbol{P}_A\boldsymbol{P}_A^{\dagger}\boldsymbol{P}_{\mathrm{FD}}^{\star}\|_{\mathrm{F}}}.$$

Finally, when we introduce $\boldsymbol{P}_D^{\star'} = \boldsymbol{P}_A^{\dagger}\boldsymbol{P}_{\mathrm{FD}}^{\star}$ for the optimal unconstrained digital precoder, we have

$$\boldsymbol{P}_D^{\star} = \frac{\sqrt{P_{\mathrm{E}}}}{\|\boldsymbol{P}_A\boldsymbol{P}_D^{\star'}\|_{\mathrm{F}}}\boldsymbol{P}_D^{\star'}.$$

We conclude that the solution for $\boldsymbol{P}_D$ in, e.g., [12, 34, 37, 82] is optimal w.r.t. to minimizing the objective function.

In the case when the consumed power P_{C} is constrained, we have the following system of equations

$$\begin{cases} \boldsymbol{p}_{\mathrm{D}}^{\star} = \left(\boldsymbol{K}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{K}_{\mathrm{A}} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{K}_{\mathrm{A}}^{\mathrm{H}}\boldsymbol{p}_{\mathrm{FD}}^{\star} \\ \|\boldsymbol{p}_{\mathrm{D}}^{\star}\|_2^2 = P_{\mathrm{C}} \\ \lambda > 0 \end{cases}. \tag{6.8}$$

A closed-form solution for λ is not available. However, we observe that $\|\boldsymbol{p}_{\mathrm{D}}^{\star}\|_2^2$ is monotonically decreasing with λ. Consequently, (6.8) may be solved, e.g., by means of the bisection method.

6.2.2. Decomposition of the Unconstrained Precoding Matrix for SPI-HBF

In the following, we show how to use the aforementioned algorithms, initially developed for FI-HBF transceivers, in the SPI-HBF scenario.

Let's extend to the notation from the previous subsection (5.7) with

$$\boldsymbol{P}_{\mathrm{FD}}^{\star} = \left[\boldsymbol{P}_{\mathrm{FD}}^{\star,(1),\mathrm{T}}, \ldots, \boldsymbol{P}_{\mathrm{FD}}^{\star,(S_{\mathrm{t}}),\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{C}^{N_{\mathrm{t}}\times N_{\mathrm{s}}}.$$

Then, we rewrite $d_0(\boldsymbol{P}_A, \boldsymbol{P}_D; \boldsymbol{P}_{\mathrm{FD}}^{\star})$ in terms of $\boldsymbol{P}_A^{(k)}$ $\boldsymbol{P}_D^{(k)}$

$$\begin{aligned} d_0\left(\left[\boldsymbol{P}_A^{(k)}\right]_{k=1}^{S_\mathrm{t}}, \left[\boldsymbol{P}_D^{(k)}\right]_{k=1}^{S_\mathrm{t}}; \left[\boldsymbol{P}_{\mathrm{FD}}^{\star,(k)}\right]_{k=1}^{S_\mathrm{t}}\right) = \\ \left\| \begin{bmatrix} \boldsymbol{P}_{\mathrm{FD}}^{\star,(1)} \\ \vdots \\ \boldsymbol{P}_{\mathrm{FD}}^{\star,(S_\mathrm{t})} \end{bmatrix} - \begin{bmatrix} \boldsymbol{P}_A^{(1)}\boldsymbol{P}_D^{(1)} \\ \vdots \\ \boldsymbol{P}_A^{(S_\mathrm{t})}\boldsymbol{P}_D^{(S_\mathrm{t})} \end{bmatrix} \right\|_{\mathrm{F}}^2 = \\ \sum_{k=1}^{S_\mathrm{t}} \|\boldsymbol{P}_{\mathrm{FD}}^{\star,(k)} - \boldsymbol{P}_A^{(k)}\boldsymbol{P}_D^{(k)}\|_{\mathrm{F}}^2. \end{aligned} \tag{6.9}$$

Consequently, we write the optimization problem (6.2) as

$$\begin{aligned} \left\{\left[\boldsymbol{P}_A^{(k),\star,0}\right]_{k=1}^{S_\mathrm{t}}, \left[\boldsymbol{P}_D^{(k),\star,0}\right]_{k=1}^{S_\mathrm{t}}\right\} = \\ \underset{\left[\boldsymbol{P}_A^{(k)}\right]_{k=1}^{S_\mathrm{t}}, \left[\boldsymbol{P}_D^{(k)}\right]_{k=1}^{S_\mathrm{t}}}{\arg\min} \sum_{k=1}^{S_\mathrm{t}} \|\boldsymbol{P}_{\mathrm{FD}}^{\star,(k)} - \boldsymbol{P}_A^{(k)}\boldsymbol{P}_D^{(k)}\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad \sum_{k=1}^{S_\mathrm{t}} \|\boldsymbol{\Gamma}\boldsymbol{P}_D^{(k)}\|_{\mathrm{F}}^2 \leq P_{\mathrm{TX}} \\ \boldsymbol{P}_A^{(k)} \in \mathcal{P}_{\mathcal{A}} \quad \forall k \in \{1, \ldots, S_\mathrm{t}\}. \end{aligned} \tag{6.10}$$

We observe the following:

- The update of the analog precoders can be performed independently for each subarray k as

$$\begin{aligned} \min_{\left[\boldsymbol{P}_A^{(k)}\right]_{k=1}^{S_\mathrm{t}} \in \mathcal{P}_{\mathcal{A}}} \sum_{k=1}^{S_\mathrm{t}} \|\boldsymbol{P}_{\mathrm{FD}}^{\star,(k)} - \boldsymbol{P}_A^{(k)}\boldsymbol{P}_D^{(k)}\|_{\mathrm{F}}^2 = \\ \sum_{k=1}^{S_\mathrm{t}} \min_{\boldsymbol{P}_A^{(k)} \in \mathcal{P}_{\mathcal{A}}} \|\boldsymbol{P}_{\mathrm{FD}}^{\star,(k)} - \boldsymbol{P}_A^{(k)}\boldsymbol{P}_D^{(k)}\|_{\mathrm{F}}^2 \end{aligned}$$

- The power budget constraint introduces dependencies between the digital precoding matrices $\boldsymbol{P}_D^{(k)}$. Therefore, the update of the entire $\boldsymbol{P}_D$ matrix should, in general, be performed as in the FI-HBF case.

Let's comment more extensively on the second point. This statement holds for the case, when the power constraint is taken into account in each alternating step. Note that in, e.g., [12,22,34] the projection of $\boldsymbol{P}_D$ onto the feasible set takes effect only *after*

the algorithm converges. In such case, the optimization w.r.t. $\boldsymbol{P}_D$ is unconstrained and can be also parallelized into S_t independent optimization problems

$$\min_{\left[\boldsymbol{P}_D^{(k)}\right]_{k=1}^{S_\mathrm{t}}} \sum_{k=1}^{S_\mathrm{t}} \|\boldsymbol{P}_{\mathrm{FD}}^{\star,(k)} - \boldsymbol{P}_A^{(k)}\boldsymbol{P}_D^{(k)}\|_\mathrm{F}^2 = \sum_{k=1}^{S_\mathrm{t}} \min_{\boldsymbol{P}_D^{(k)}} \|\boldsymbol{P}_{\mathrm{FD}}^{\star,(k)} - \boldsymbol{P}_A^{(k)}\boldsymbol{P}_D^{(k)}\|_\mathrm{F}^2. \tag{6.11}$$

We summarize the decomposition procedure for SPI-HBF in Algorithm 4.

6.3. Precoding With the Multiuser Perspective

The analysis of the performance bounds for SPI-HBF suggests another perspective in the discussion about subarray systems. The equivalence between SPI-HBF and a cooperative multiuser system, where transceivers have limited multiplexing abilities, has been shown. We utilize this novel perspective to propose heuristic precoding schemes which emerge from the field of research related to multiuser precoding.

In the following, we propose to design the analog and digital precoding in two separate steps.

First, we treat all the subarrays at the Tx as non-cooperative agents in a multiuser system, while the Rx subarrays can cooperate in order to jointly decode information. Moreover, the Tx and Rx agents have limited multiplexing abilities—each can transmit (receive) less independent data streams than the number of available antennas. For such scenario, we develop transmit and receive strategies—the precoding (combining) matrices $\boldsymbol{P}_A^{(k)}$ $\left(\boldsymbol{G}_A^{(l)}\right)$.

In the second step, we introduce cooperation between the transmitters. For sake of maximizing the performance of the system, we design digital precoding and combining matrices ($\boldsymbol{P}_D$ and $\boldsymbol{G}_D$).

This procedure reflects itself in reformulation of the general rate optimization problem which reads

$$C_{\boldsymbol{H}} = \max_{\boldsymbol{P}_D,\boldsymbol{P}_A,\boldsymbol{G}_D,\boldsymbol{G}_A} R_{\boldsymbol{H}} \text{ s.t. } \|\boldsymbol{\Gamma}\boldsymbol{P}_D\|_\mathrm{F}^2 \leq P_\mathrm{TX},\ \boldsymbol{P}_A \in \mathcal{P}_\mathcal{A} \tag{6.12}$$

into one step of alternating optimization between the analog precoders (combiners) $\boldsymbol{P}_A$ ($\boldsymbol{G}_A$) and the digital ones: $\boldsymbol{P}_D$ and $\boldsymbol{G}_D$. First, we initialize $\boldsymbol{P}_D = \boldsymbol{I}$ and $\boldsymbol{G}_D = \boldsymbol{I}$

and solve the optimization problem with respect to the analog stage matrices

$$\{\boldsymbol{P}_A^\star, \boldsymbol{G}_A^\star\} = \underset{\boldsymbol{P}_A, \boldsymbol{G}_A}{\arg\max} \log_2 \det\left(\boldsymbol{I} + \boldsymbol{R}_\eta^{-1} \boldsymbol{G}_A^{\mathrm{H}} \boldsymbol{H} \boldsymbol{P}_A \boldsymbol{P}_A^{\mathrm{H}} \boldsymbol{H}^{\mathrm{H}} \boldsymbol{G}_A\right)$$
$$\text{s.t. } \boldsymbol{P}_A \in \mathcal{P}_{\mathcal{A}},\ \boldsymbol{G}_A \in \mathcal{G}_{\mathcal{A}} \tag{6.13}$$

and subsequently with respect to the digital stage matrices

$$\{\boldsymbol{P}_D^\star, \boldsymbol{G}_D^\star\} = \underset{\boldsymbol{P}_D, \boldsymbol{G}_D}{\arg\max} \log_2 \det\left(\boldsymbol{I} + \boldsymbol{R}_{\eta,G}^{-1} \boldsymbol{G}_D^{\mathrm{H}} \boldsymbol{H}_{\mathrm{A}}^\star \boldsymbol{P}_D \boldsymbol{P}_D^{\mathrm{H}} \boldsymbol{H}_{\mathrm{A}}^{\star,\mathrm{H}} \boldsymbol{G}_D\right)$$
$$\text{s.t. } \|\boldsymbol{\Gamma} \boldsymbol{P}_D\|_{\mathrm{F}}^2 \leq P_{\mathrm{TX}}. \tag{6.14}$$

where $\boldsymbol{H}_{\mathrm{A}}^\star = \boldsymbol{G}_A^{\star,\mathrm{H}} \boldsymbol{H} \boldsymbol{P}_A^\star$ and $\boldsymbol{R}_{\eta,G} = \boldsymbol{G}_A^{\star,\mathrm{H}} \boldsymbol{R}_\eta \boldsymbol{G}_A^\star$. We denote the rate achievable with such strategy as

$$R_{\boldsymbol{H},\mathrm{MU}} = \log_2 \det(\boldsymbol{I} + \boldsymbol{R}_\eta^{\star,-1} \boldsymbol{G}_D^{\star\mathrm{H}} \boldsymbol{G}_A^{\star\mathrm{H}} \boldsymbol{H} \boldsymbol{P}_A^\star \boldsymbol{P}_D^\star \boldsymbol{P}_D^{\star\mathrm{H}} \boldsymbol{P}_A^{\star\mathrm{H}} \boldsymbol{H}^{\mathrm{H}} \boldsymbol{G}_A^{\star\mathrm{H}} \boldsymbol{G}_D^{\star\mathrm{H}}) \tag{6.15}$$

where $\boldsymbol{R}_\eta^\star = \boldsymbol{G}_D^{\star\mathrm{H}} \boldsymbol{G}_A^{\star\mathrm{H}} \boldsymbol{R}_\eta \boldsymbol{G}_A^\star \boldsymbol{G}_D^\star$.

Clearly, separating the optimization problem variables results in a non-optimal solution $R_{\boldsymbol{H},\mathrm{MU}} \leq C_{\boldsymbol{H}}$. However, attaining the global optimum of (6.12) is not guaranteed with methods available in the literature[1]; not only the non-convex unity-modulus constraint, but also the coupling of the optimization variables is problematic.

We show that by separating the problem as in (6.13) and (6.14), we get two problems which are easier to solve. We tackle the first problem by adapting algorithms developed for multiuser linear precoding scenarios. The second problem is to determine the capacity of a point-to-point MIMO setup with deterministic channel which is solved, e.g., in [24, 71].

We note that there exists a substantial number of linear precoding methods for multiuser MIMO. A significant portion of them can be adopted for solving the problem in question. In this work, we develop an algorithm inspired by the LISA algorithm which has been shown in [26] for a fully digital scenario. We motivate choosing this approach with high flexibility of the algorithm and that it converges, for particular subarray setups, to other algorithms which were shown to attain very good results, namely the Hybrid-LISA algorithm in [76] and the precoding presented in [18].

In the following, we explain the details of the algorithm. We note that it has been previously presented in [36]. In this work, we provide more details and discussion on the proposed method.

[1] A local optimum can be determined, e.g., by means of the ADMM algorithm [75], although convergence is not guaranteed.

6.3.1. The SGSGP Algorithm Description

In (5.5), we introduced a decomposition of $\boldsymbol{H}$ into $S_\mathrm{t}S_\mathrm{r}$ submatrices $\boldsymbol{H}_{l,k}$, each corresponding to a channel transfer matrix between the lth Rx subarray and kth Tx subarray. The row- and null-spaces of the submatrices govern how information may propagate between the subarrays. For example, let $\boldsymbol{p} \in \text{rowspace}(\boldsymbol{H}_{l,k})$ and the transmit vector $\boldsymbol{x} = \boldsymbol{p}s$ be transmitted by the kth subarray. The symbol s might be then recovered by each subarray l, for which $\boldsymbol{p} \notin \text{nullspace}(\boldsymbol{H}_{l,k})$.

The subarray l may then decide to either receive the symbol by combining the received vector $\boldsymbol{H}_{l,k}\boldsymbol{p}s$ with a vector $\boldsymbol{g}^\mathrm{T} \notin \text{nullspace}(\boldsymbol{H}_{l,k}\boldsymbol{p})$, such that $\hat{s} = \boldsymbol{g}^\mathrm{T}\boldsymbol{H}_{l,k}\boldsymbol{p}s \neq 0$ or neglect it by setting $\boldsymbol{g}^\mathrm{T} \in \text{nullspace}(\boldsymbol{H}_{l,k}\boldsymbol{p})$. In our algorithm we keep these facts in mind. We iteratively assign streams to a pair of transmit and receive subarrays—we assign to them precoding vectors $\boldsymbol{p}$ that lie in the rowspace of the corresponding channel. We control the interference to the previously assigned symbols by separating the precoding vectors from the previous ones by slicing the rowspace into subspaces corresponding to different symbols. At the receiver side, we may avoid the interference from previously allocated streams if we receive in the nullspace of the space occupied by them. In the following, we formalize this approach, which we call Spatial Greedy Subarray Hybrid Precoding (SGSHP).

In the first step of the algorithm, we search for a pair of subarrays (l_1, k_1) that maximizes the gain of the allocated stream

$$\begin{aligned}\{l_1, k_1, \boldsymbol{g}_1, \boldsymbol{p}_1\} = \arg\max_{k,l,\boldsymbol{g},\boldsymbol{p}} &|\boldsymbol{g}^\mathrm{H}\boldsymbol{H}_{l,k}\boldsymbol{p}| \\ \text{s.t. } &\boldsymbol{g} \in \mathcal{G}_\mathcal{A},\ \boldsymbol{p} \in \mathcal{P}_\mathcal{A}.\end{aligned} \tag{6.16}$$

We notice that for any $\boldsymbol{g}$, the objective is maximized if

$$\boldsymbol{p}(\boldsymbol{g}) = \xi^{(k)}_{A,\mathrm{TX}} \exp(j\angle(\boldsymbol{H}^\mathrm{H}_{l,k}\boldsymbol{g})). \tag{6.17}$$

As

$$\begin{aligned}|\boldsymbol{g}^\mathrm{H}\boldsymbol{H}_{l,k}\xi^{(k)}_{A,\mathrm{TX}} \exp(j\angle({\boldsymbol{H}_{l,k}}^\mathrm{H}\boldsymbol{g}))| &= \xi^{(k)}_{A,\mathrm{TX}} \sum_i \left|\left[\boldsymbol{g}^\mathrm{H}\boldsymbol{H}_{l,k}\right]_i\right| \\ &= \xi^{(k)}_{A,\mathrm{TX}} \|\boldsymbol{g}^\mathrm{H}\boldsymbol{H}_{l,k}\|_1\end{aligned} \tag{6.18}$$

we rewrite the problem only as a function of $\boldsymbol{g}$ as

$$\begin{aligned}\{k_1, l_1, \boldsymbol{g}_1\} = \arg\max_{k,l,\boldsymbol{g}} &\xi^{(k)}_{A,\mathrm{TX}} \|\boldsymbol{g}^\mathrm{H}\boldsymbol{H}_{l,k}\|_1 \\ \text{s.t. } &\boldsymbol{g} \in \mathcal{G}_\mathcal{A}.\end{aligned} \tag{6.19}$$

This problem qualitatively corresponds to the L1-norm principal component analysis (PCA-L1) problem formulated in [40] with the additional constraint on the entries of $\boldsymbol{g}$. We therefore propose to relax the constraint from (6.19) to the l2 constraint $\|\boldsymbol{g}'\|_2 \leq \sqrt{N_{\mathrm{r}}^k}\xi_{A,\mathrm{RX}}^{(l)}$ and consequently solve a following problem

$$\begin{aligned}\{k_1, l_1, \boldsymbol{g}_1'\} = \arg\max_{k,l,\boldsymbol{g}'} \xi_{A,\mathrm{TX}}^{(k)} \|\boldsymbol{g}'^{\mathrm{H}} \boldsymbol{H}_{l,k}\|_1 \\ \text{s.t.} \quad \|\boldsymbol{g}'\|_2 \leq \sqrt{N_{\mathrm{r}}^k}\xi_{A,\mathrm{RX}}^{(l)}.\end{aligned} \tag{6.20}$$

utilizing the greedy optimization algorithm from [40][2]. Afterwards, we project the vector $\boldsymbol{g}_1'$ back into the nearest, in the Euclidean sense, element of $\mathcal{G}_{\mathcal{A}}$, i.e., $[\boldsymbol{g}_1]_n = \frac{[\boldsymbol{g}_1']_n}{|[\boldsymbol{g}_1']_n|}\xi_{A,\mathrm{RX}}^{(l)}$ [22]. Finally, we determine $\boldsymbol{p}_1$ as

$$\boldsymbol{p}_1 = \xi_{A,\mathrm{TX}}^{(k)} \exp(j\angle(\boldsymbol{H}_{l,k}^{\mathrm{H}} \boldsymbol{g}_1)). \tag{6.21}$$

Starting from the second step of the algorithm, we limit the interference caused to the already allocated streams by controlling the subspace in which the next stream is allocated. To this end, we introduce a set of projection matrices $\boldsymbol{\Xi}_k^{(i)}$ and choose the precoding vector within the subspace spanned by the columns of $\boldsymbol{\Xi}_k^{(i)} \boldsymbol{H}_{:,k}^{\mathrm{T}}$. Moreover, we limit the interference from the already allocated streams by receiving only within the subspace $(\boldsymbol{\Pi}_l^{(i)} \boldsymbol{H}_{l,:})^{\mathrm{T}}$, where $\boldsymbol{\Pi}_l^{(i)}$ is a projection matrix (construction of which we discuss further). For this sake, we write the precoding and combining vectors as $\boldsymbol{p} = \boldsymbol{\Xi}_k^{(i)} \boldsymbol{m}$, $\boldsymbol{g} = \boldsymbol{\Pi}_l^{(i)} \boldsymbol{t}$ and formulate the following optimization problem

$$\begin{aligned}\{k_i, l_i, \boldsymbol{t}_i, \boldsymbol{m}_i\} = \arg\max_{k,l,\boldsymbol{t},\boldsymbol{m}} |\boldsymbol{t}^{\mathrm{H}} \boldsymbol{\Pi}_l^{(i)} \boldsymbol{H}_{l,k} \boldsymbol{\Xi}_k^{(i)} \boldsymbol{m}| \\ \text{s.t.} \quad \boldsymbol{t}^{\mathrm{H}} \boldsymbol{\Pi}_l^{(i)} \in \mathcal{G}_{\mathcal{A}}, \ \boldsymbol{\Xi}_k^{(i)} \boldsymbol{m} \in \mathcal{P}_{\mathcal{A}},\end{aligned} \tag{6.22}$$

The introduction of the projection matrices into the constraints have complicated the solution comparing to (6.16). The problem is non-convex (the function to be maximized is convex and the constraints are non-convex) and not tractable analytically. We propose to relax both constraints and project the unconstrained solution onto the feasible set. For this sake, we first solve

$$\begin{aligned}\{k_i', l_i', \boldsymbol{t}_i', \boldsymbol{m}_i'\} = \arg\max_{k,l,\boldsymbol{t},\boldsymbol{m}} |\boldsymbol{t}^{\mathrm{H}} \boldsymbol{\Pi}_l^{(i)} \boldsymbol{H}_{l,k} \boldsymbol{\Xi}_k^{(i)} \boldsymbol{m}| \\ \text{s.t.} \quad \|\boldsymbol{t}\|_2 = 1, \|\boldsymbol{m}\|_2 = 1.\end{aligned} \tag{6.23}$$

[2] We note that a simple extension to the algorithm, where in each iterative step the solution would have been projected into $\mathcal{G}_{\mathcal{A}}$ would lack convergence properties proven in [40].

This problem has a closed form solution—$\boldsymbol{t}'_i, \boldsymbol{m}'_i$ are left and right singular vectors of $\boldsymbol{\Pi}_l^{(i)} \boldsymbol{H}_{l,k} \boldsymbol{\Xi}_k^{(i)}$ corresponding to the largest singular value of all the $S_\mathrm{t} S_\mathrm{r}$ possible matrices. As our actual interest are the precoding and combining vectors $\boldsymbol{p}_i$ and $\boldsymbol{q}_i$ we write $\boldsymbol{p}'_i = \boldsymbol{\Xi}_{k'_i}^{(i)} \boldsymbol{m}'_i,\ \boldsymbol{g}'_i = \boldsymbol{\Pi}_{l'_i}^{(i)} \boldsymbol{t}'_i$ and project those onto the closest (in the Euclidean sense) elements of the feasible sets

$$\begin{aligned} \boldsymbol{p}_i := [\boldsymbol{p}_i]_n &= \frac{[\boldsymbol{p}'_i]_n}{|[\boldsymbol{p}'_i]_n|} \xi_{A,\mathrm{TX}}^{(k)} \\ \boldsymbol{g}_i := [\boldsymbol{g}_i]_n &= \frac{[\boldsymbol{g}'_i]_n}{|[\boldsymbol{g}'_i]_n|} \xi_{A,\mathrm{RX}}^{(l)} \end{aligned} \tag{6.24}$$

and assign $k'_i := k,\ l'_i := l$.

In the following, we discuss in detail the possible construction of the projection matrices.

We note that complete inter-stream interference mitigation means enforcing

$$\forall i \neq j \implies \boldsymbol{g}_i^\mathrm{H} \boldsymbol{H}_{l_i,k_j} \boldsymbol{p}_j = 0$$

This would be possible by updating the projectors in the i-th step as follows

$$\begin{aligned} \forall (l,k) \in \{1,\dots,S_\mathrm{r}\} \times \{1,\dots,S_\mathrm{t}\} & \\ \boldsymbol{t}_k^\mathrm{H} &= \boldsymbol{g}_i^\mathrm{H} \boldsymbol{H}_{l_i,k} \boldsymbol{\Xi}_k^{(i)} \\ \boldsymbol{r}_l &= \boldsymbol{\Pi}_l^{(i)} \boldsymbol{H}_{l,k_i} \boldsymbol{p}_i \\ \boldsymbol{\Xi}_k^{(1)} = \boldsymbol{I}\ \forall k,\ \boldsymbol{\Xi}_k^{(i+1)} &= \boldsymbol{\Xi}_k^{(i)} - \boldsymbol{t}_k (\boldsymbol{t}_k^\mathrm{H} \boldsymbol{t}_k)^{-1} \boldsymbol{t}_k^\mathrm{H} \\ &= \boldsymbol{\Xi}_k^{(i)} - \frac{\boldsymbol{t}_k \boldsymbol{t}_k^\mathrm{H}}{\|\boldsymbol{t}_k\|_2^2}, \\ \boldsymbol{\Pi}_l^{(1)} = \boldsymbol{I}\ \forall l,\ \boldsymbol{\Pi}_l^{(i+1)} &= \boldsymbol{\Pi}_l^{(i)} - \frac{\boldsymbol{r}_l \boldsymbol{r}_l^\mathrm{H}}{\|\boldsymbol{r}_l\|_2^2}. \end{aligned} \tag{6.25}$$

and solving (6.22) optimally. The projectors orthogonally project out the space corresponding to the currently allocated stream, both at the Rx and Tx.

Such conservative zero-forcing approach can often be a wrong choice as it significantly reduces the degrees of freedom (in this case, the remaining subspace) for allocating the consecutive streams. This is expected to be especially pronounced in the mmWave frequencies, where the channel is in general low-rank.

On the other hand, a significant portion of interference is suppressed given that in

the i-th step only $\boldsymbol{\Xi}_{k_i}^{(i)}$ and $\boldsymbol{\Pi}_{l_i}^{(i)}$ are updated, which can be written as following

$$\begin{aligned}
\boldsymbol{t}^{\mathrm{H}} &= \boldsymbol{g}_i^{\mathrm{H}} \boldsymbol{H}_{l_i,k_i} \boldsymbol{\Xi}_{k_i}^{(i)} = \sigma_t \boldsymbol{p}_i^{\mathrm{H}} \\
\boldsymbol{r}^{\mathrm{H}} &= \boldsymbol{\Xi}_{l_i}^{(i)} \boldsymbol{H}_{l_i,k_i} \boldsymbol{p}_i = \sigma_r \boldsymbol{g}_i \\
\boldsymbol{\Xi}_k^{(1)} = \boldsymbol{I}\ \forall k,\ \boldsymbol{\Xi}_{k_i}^{(i+1)} &= \boldsymbol{\Xi}_{k_i}^{(i)} - \boldsymbol{t}(\boldsymbol{t}^{\mathrm{H}}\boldsymbol{t})^{-1}\boldsymbol{t}^{\mathrm{H}} \\
&= \boldsymbol{\Xi}_{k_i}^{(i)} - \boldsymbol{p}_i \boldsymbol{p}_i^{\mathrm{H}} \\
\boldsymbol{\Pi}_l^{(1)} = \boldsymbol{I}\ \forall l,\ \boldsymbol{\Pi}_{l_i}^{(i+1)} &= \boldsymbol{\Pi}_{l_i}^{(i)} - \boldsymbol{r}(\boldsymbol{r}^{\mathrm{H}}\boldsymbol{r})^{-1}\boldsymbol{r}^{\mathrm{H}} \\
&= \boldsymbol{\Pi}_{l_i}^{(i)} - \boldsymbol{g}_i \boldsymbol{g}_i^{\mathrm{H}}.
\end{aligned} \tag{6.26}$$

It can be verified that with such construction, following is enforced

$$j > i \wedge l_j = l_i \implies \boldsymbol{g}_j^{\mathrm{H}} \boldsymbol{H}_{l_j,k_i} \boldsymbol{p}_i = 0 \tag{6.27}$$

$$j > i \wedge k_j = k_i \implies \boldsymbol{g}_i^{\mathrm{H}} \boldsymbol{H}_{l_i,k_j} \boldsymbol{p}_j = 0 \tag{6.28}$$

In other words, the already allocated streams do not *interfere* other streams allocated later to the same Rx subarray (6.27) and do not *experience interference* from streams later allocated to the same Tx subarray (6.28).

The details of the solutions are presented in Algorithm 5. We note that the construction of the projection matrices has not been specified, as the best choice can vary depending on the particular setup.

6.4. Low-SNR precoding

In the low-SNR region, noise is the factor limiting a MIMO link. Therefore, the optimal strategy is to trade multiplexing gain for power gain. Consequently, the optimal number of streams becomes lower as the SNR decreases.

We exploit this fact in our proposal of a precoding strategy for the low-SNR region. We realize that the number of available spatial degrees-of-freedom (given usual conditions, equal to the number of RF chains) in SPI-HBF may exceed the demand. In the following, we formulate a theorem showing that under certain conditions, a SPI-HBF architecture with more RF chains can exactly mimic a FI-HBF with less RF chains. As a consequence of this fact, we may reuse well established FI-HBF algorithms for SPI-HBF beamforming and achieve good performance, given the low number of FI-HBF RF chains is sufficient compared to the stream number requirement.

Theorem 1. Any precoding realized by a FI-HBF Tx with N_{RF} RF chains may be realized by a SPI-HBF Tx with equal number of antennas given $N_{\mathrm{RF}} \leq \min_k\{N_{\mathrm{t}}^{\mathrm{RF},k}\}$, where $N_{\mathrm{t}}^{\mathrm{RF},k}$ is the number of RF chains of the kth SPI-HBF subarray.

Proof. In the following, the FI-HBF transmitter is equipped with N'_{RF} RF chains and ${N_{\mathrm{t}}}'$ antennas. The SPI-HBF is equipped with N''_{RF} RF chains and ${N_{\mathrm{t}}}''$ antennas.

We denote the FI-HBF analog precoding matrix as $\boldsymbol{P}'_A$ and the digital precoding matrix with $\boldsymbol{P}'_D$. Similarly, let the SPI-HBF analog and digital precoding matrix read $\boldsymbol{P}''_A$ and $\boldsymbol{P}''_D$, respectively. In the following, we first provide a proof considering a special case for illustration purposes. The proof of the general case follows afterwards.

Case 1. In the first case, we consider a SPI-HBF setup with K subarrays, each having equal number of RF chains

$$N_{\mathrm{t}}^{\mathrm{RF},1} = \cdots = N_{\mathrm{t}}^{\mathrm{RF},K} = \frac{N''_{\mathrm{RF}}}{K} \tag{6.29}$$

and the FI-HBF has $N'_{\mathrm{RF}} = \frac{N''_{\mathrm{RF}}}{K}$ RF chains in total such that all K inequalities $N'_{\mathrm{RF}} \leq N_{\mathrm{t}}^{\mathrm{RF},k}\ \forall k$ are met with equality $N'_{\mathrm{RF}} = N_{\mathrm{t}}^{\mathrm{RF},k}\ \forall k$. The number of antennas at the subarrays (N_{t}^k for the kth subarray) may differ and only the following must hold

$$\sum_{k=1}^{K} N_{\mathrm{t}}^k = {N_{\mathrm{t}}}'' = {N_{\mathrm{t}}}' \tag{6.30}$$

Subsequently, we write the FI-HBF analog precoding matrix $\boldsymbol{P}'_A$ as a concatenation of K matrices $\boldsymbol{P}_A^{(1)} \in \mathbb{C}^{N_{\mathrm{t}}^1 \times N'_{\mathrm{RF}}}, \ldots, \boldsymbol{P}_A^{(K)} \in \mathbb{C}^{N_{\mathrm{t}}^K \times N'_{\mathrm{RF}}}$

$$\boldsymbol{P}'_A = \begin{bmatrix} \boldsymbol{P}_A^{(1)} \\ \vdots \\ \boldsymbol{P}_A^{(K)} \end{bmatrix}$$

Then if we construct the SPI-HBF analog and digital precoding matrices as following

$$\boldsymbol{P}''_A = \begin{bmatrix} s_1^1 \boldsymbol{P}_A^{(1)} & & \\ & \ddots & \\ & & s_1^K \boldsymbol{P}_A^{(K)} \end{bmatrix}$$

$$\boldsymbol{P}''_D = \begin{bmatrix} \frac{1}{s_1^1} \boldsymbol{P}'_D \\ \vdots \\ \frac{1}{s_1^K} \boldsymbol{P}'_D \end{bmatrix}$$

we have $\boldsymbol{P}_A''\boldsymbol{P}_D'' = \boldsymbol{P}_A'\boldsymbol{P}_D'$, therefore the equality of the FI-HBF and SPI-HBF effective precoding matrix. The scaling with $s_1^k = \sqrt{\frac{N_\mathrm{t}'}{N_\mathrm{t}^k}}$ ensures that $\boldsymbol{P}_A'' \in \mathcal{P}_\mathcal{A}$.

Case 2. Here, a more general case is considered. The inequalities no longer need to be met with equality, i.e., we let the subarrays to have more RF chains available than the FI-HBF. We then complement the presented construction with introduction of "dummy" matrices $\boldsymbol{M}^k \in \mathbb{C}^{N_\mathrm{t}^k \times (N_\mathrm{t}^{\mathrm{RF},k} - N_\mathrm{RF}')}$ such that $\left|[\boldsymbol{M}^k]_{i,j}\right| = \xi_{A,\mathrm{TX}}^{(k)}$ and write the SPI-HBF precoding matrix as

$$\boldsymbol{P}_A'' = \begin{bmatrix} \left[s_2^1 \boldsymbol{P}_A^{(1)}, \boldsymbol{M}^1\right] & & \\ & \ddots & \\ & & \left[s_2^K \boldsymbol{P}_A^{(K)}, \boldsymbol{M}^K\right] \end{bmatrix},$$

$$\boldsymbol{P}_D'' = \begin{bmatrix} \frac{1}{s_2^1}\boldsymbol{P}_D' \\ \mathbf{0}_{(N_\mathrm{t}^{\mathrm{RF},1} - N_\mathrm{RF}') \times N_\mathrm{s}'} \\ \vdots \\ \frac{1}{s_2^K}\boldsymbol{P}_D' \\ \mathbf{0}_{(N_\mathrm{t}^{\mathrm{RF},K} - N_\mathrm{RF}') \times N_\mathrm{s}'} \end{bmatrix}$$

and achieve equality of the FI-HBF and SPI-HBF effective precoding matrices $\boldsymbol{P}_A''\boldsymbol{P}_D'' = \boldsymbol{P}_A'\boldsymbol{P}_D'$. The scaling with $s_2^k = \sqrt{\frac{N_\mathrm{t}' N_\mathrm{RF}'}{N_\mathrm{t}^k N_\mathrm{t}^{\mathrm{RF},k}}}$ ensures that $\boldsymbol{P}_A'' \in \mathcal{P}_\mathcal{A}$.

□

The proof shows that the precoding of a FI-HBF Tx with N_RF' RF chains may be replicated by a SPI-HBF Tx with equal number of antennas given $N_\mathrm{t}^{\mathrm{RF},k} \geq N_\mathrm{RF}' \; \forall k$. We note that the maximum number of streams that may be encoded is limited by N_RF', a number lower than the number of SPI-HBF RF chains (lower bounded by $S_\mathrm{t} N_\mathrm{RF}'$). Therefore, the approach is reasonable in the low-SNR region, where the optimal number of streams is low (in the range of 1 to $\min\{N_\mathrm{t}^{\mathrm{RF},1}, \ldots, N_\mathrm{t}^{\mathrm{RF},S_\mathrm{t}}\}$). We summarize the procedure in Algorithm 6.

6.5. Realizing Fully-Digital Beamforming with SPI-HBF

From Theorem 1 we yield that a SPI architecture can realize the same precoding as a FI architecture with $\min\{N_\mathrm{t}^{\mathrm{RF},k}\}$ of RF chains. Moreover, it has been shown in [67]

that for FI-HBF, the full digital beamforming can be perfectly reconstructed given $2N_{\mathrm{s}} \leq N_{\mathrm{t}}^{\mathrm{RF}}$. Consequently, we formulate a following theorem

Theorem 2. A SPI-HBF precoding architecture can realize fully-digital beamforming given $2N_{\mathrm{s}} \leq \min\{N_{\mathrm{t}}^{\mathrm{RF},1}, \ldots, N_{\mathrm{t}}^{\mathrm{RF},S_{\mathrm{t}}}\}$.

Proof. The proof follows as a straightforward corollary from the Proposition 2 from [67] and Theorem 1.

From Theorem 1 we have that given $N_{\mathrm{s}} \leq \min\{N_{\mathrm{t}}^{\mathrm{RF},1}, \ldots, N_{\mathrm{t}}^{\mathrm{RF},S_{\mathrm{t}}}\}$, a SPI-HBF transmitter is equivalent to a FI-HBF transmitter with $N_{\mathrm{t}}^{\mathrm{RF}} = \min\{N_{\mathrm{t}}^{\mathrm{RF},1}, \ldots, N_{\mathrm{t}}^{\mathrm{RF},S_{\mathrm{t}}}\}$. Moreover, from Proposition 2 from [67] we have that full digital precoding can be realized given $2N_{\mathrm{s}} \leq N_{\mathrm{t}}^{\mathrm{RF}}$. Consequently, a SPI-HBF with subarrays equipped with $N_{\mathrm{t}}^{\mathrm{RF},1}, \ldots, N_{\mathrm{t}}^{\mathrm{RF},S_{\mathrm{t}}}$ RF chains can realize full digital beamforming given

$$2N_{\mathrm{s}} \leq \min\{N_{\mathrm{t}}^{\mathrm{RF},1}, \ldots, N_{\mathrm{t}}^{\mathrm{RF},S_{\mathrm{t}}}\}.$$

□

We note that this observation guarantees that in low SNR regimes, where the optimal number of streams is lower or equal to

$$\left\lfloor \frac{1}{2} \min\left\{N_{\mathrm{t}}^{\mathrm{RF},1}, \ldots, N_{\mathrm{t}}^{\mathrm{RF},S_{\mathrm{t}}}\right\} \right\rfloor,$$

the capacity is achievable with SPI-HBF.

6.6. Numerical results

In this section, we concentrate on three different aspects. The first one continues the analysis of the new upper bounds—we show in which setups the precoding algorithms perform close to the newly derived limits. The second one investigates how different power constraints affect the comparison between SPI-HBF and FI-HBF. Finally, we compare the newly derived precoding method SGSHP with other existing algorithms.

Depending on the setup considered, the novel upper bounds provide a different degree of insight into SPI-HBF performance. In cases where the analog stage at the Rx or Tx is fully connected, we expect the bound to be tight at high-SNR. This might be not anymore the case when the structure of both Rx or Tx consists of multiple subarrays—the high-SNR bound does not take into account the joint subarray configuration of Tx and Rx, only the one more affecting the rate. Additionally,

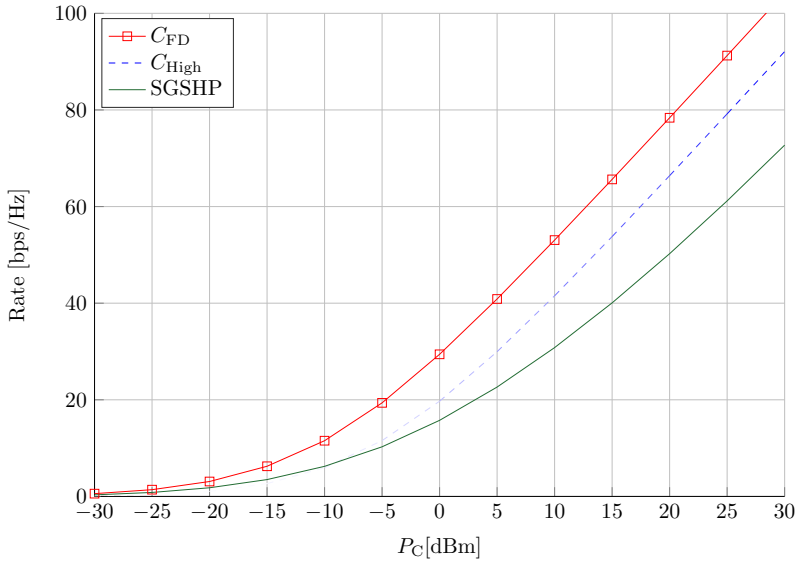

Figure 6.1.. Rate achievable with SGSHP for $S_{\mathrm{t}} = 8, S_{\mathrm{r}} = 8$, related to the novel high-SNR upper bound C_{High} and the constrained capacity C_{FD}. The antenna setup is asymmetrical, with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx.

the high-SNR upper bound should prove its usefulness with multiple Tx/Rx subarrays, when the constrained capacity is expected to be an especially loose bound. In Figs. 6.1-6.3 we show three examples where the new upper bound provides little, moderate, and substantial insight into the performance of the precoding algorithm (we use SGSHP for sake of presentation). We note that C_{FD} changes in each of the plots as P_{C} is constrained, so the emitted power budget changes with the number of transmit subarrays as in (4.7).

In Fig. 6.1 we consider 8 subarrays at the Tx and 8 subarrays at the Rx. As can be seen, the new bound is loose with respect to the SGSHP rate and does not provide much new information compared to the constrained capacity. In Fig. 6.2 we decrease the number of subarrays at the Rx to 2 and remain with 8 subarrays at the Tx. Although the bound is still not tight, it is significantly closer than the constrained capacity. Finally, in Fig. 6.3 we present results for 8 subarrays at the Tx and 1 subarray at the Rx (corresponding to a FI-HBF). Here the bound is tight for the SGSHP precoding, which is a very useful insight—it proves that for such setup the room of improvement is very limited, while the comparison with the constrained capacity suggests something contrary.

In the next step, we compare the precoding performance (for sake of presentation,

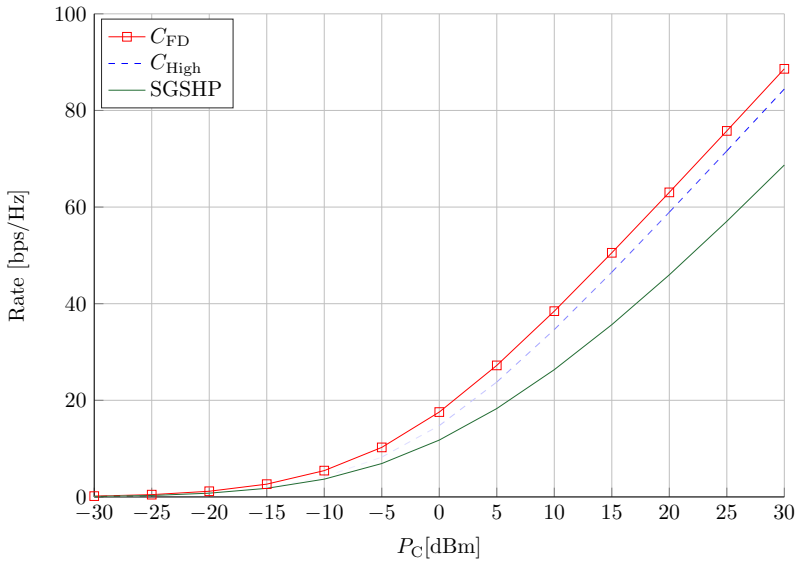

Figure 6.2.. Rate achievable with SGSHP for $S_{\mathrm{t}} = 8, S_{\mathrm{r}} = 2$, related to the novel high-SNR upper bound C_{High} and the constrained capacity C_{FD}. The antenna setup is asymmetrical, with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx.

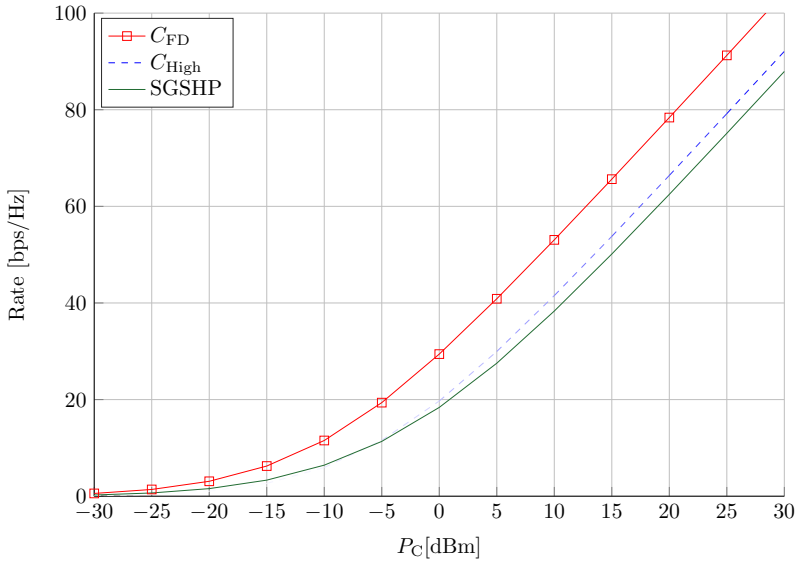

Figure 6.3.. Rate achievable with SGSHP for $S_{\mathrm{t}} = 8, S_{\mathrm{r}} = 1$, related to the novel high-SNR upper bound C_{High} and the constrained capacity C_{FD}. The antenna setup is asymmetrical, with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx.

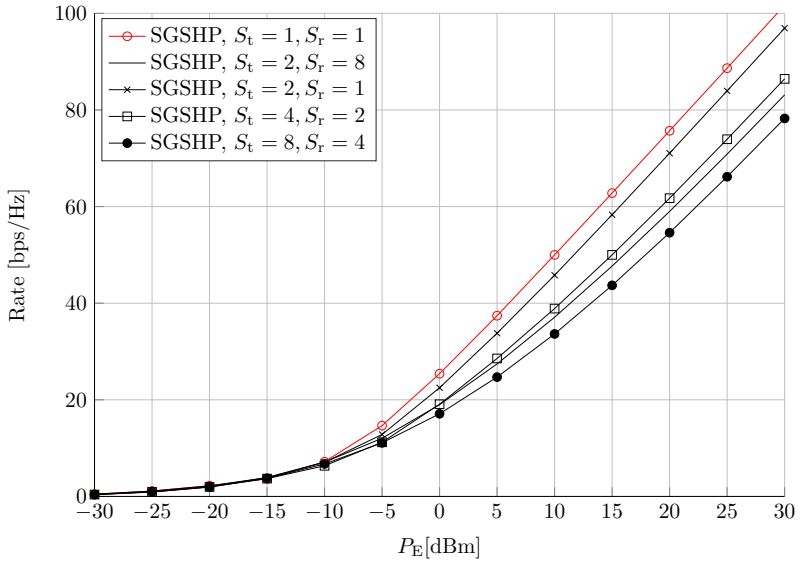

Figure 6.4.. The rate achievable with SGSHP for different subarray setups with the emitted power constraint. The antenna setup is asymmetrical, with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx.

we consider SGSHP) for different power constraints, i.e., for emitted and consumed power constraint. The latter promotes structures with multiple subarrays at the Tx—as the losses within the structure are lower, more power is emitted which in turn shifts the achievable rate curves horizontally.

In Fig. 6.4, rates achieved with different subarray setups are compared, all with equal emitted power. The results we observe are intuitive—with increasing number of subarrays, either at the Tx or the Rx, the performance drops. The inference becomes less straightforward when the *consumed* power is constrained. There, two effects counteract. On the one hand, more subarrays (either at the Tx or Rx) reduce the degrees of freedom for precoder design and thus negatively affect the performance. On the other hand, transmitters with more subarrays experience less power loss, which in turn results in more emitted power. In Fig. 6.5 we show a few examples when the performance of FI-HBF can be exceeded with SPI-HBF (for 2 Tx subarrays with FI-HBF Rx, for 4 Tx subarrays with 2 Rx subarrays, and for 8 Tx subarrays already with 4 Rx subarrays) and a case when the power gain is not sufficient to counteract the loss of degrees of freedom (for 2 Tx subarrays and 8 Rx subarrays).

In the following, we compare different precoding methods for SPI-HBF. We consider following approaches:

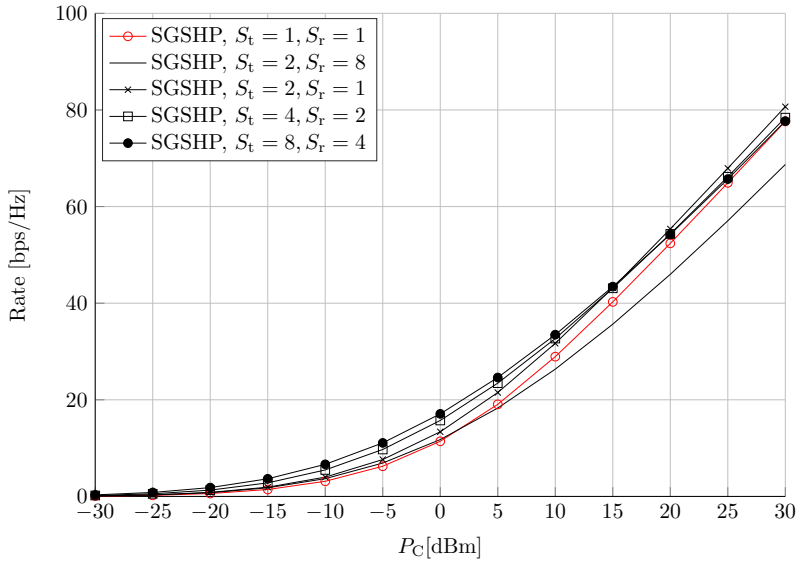

Figure 6.5.. The rate achievable with SGSHP for different subarray setups with the consumed power constraint. The antenna setup is asymmetrical, with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx.

- the OMP precoding tailored for SPI-HBF [12],
- the BCD approach tailored for SPI-HBF [22],
- the MO-AltMin algorithm [82],
- the SGSHP algorithm described in this thesis,
- the low-SNR approach described in this thesis.

The OMP, BCD, and MO-AltMin algorithms show good performance and are commonly considered as State-of-the-Art (SotA). We show when the proposed approaches outperform the SotA and when they do not provide any gains.

We focus on comparing different precoding schemes and we therefore present the results only with the emitted power constraint.

The results shown in Figs. 6.6-6.10 provide some interesting insights. One thing to notice is the performance of the introduced SGSHP scheme. In each case (also in other subarray configurations, not presented in the thesis) the performance is better or equal to the best results from the SotA—BCD or MO-AltMin. Here we stress that both BCD and SGSHP are of much lower computational complexity than the MO-AltMIN, which exhibits exponential complexity with the number of antennas.

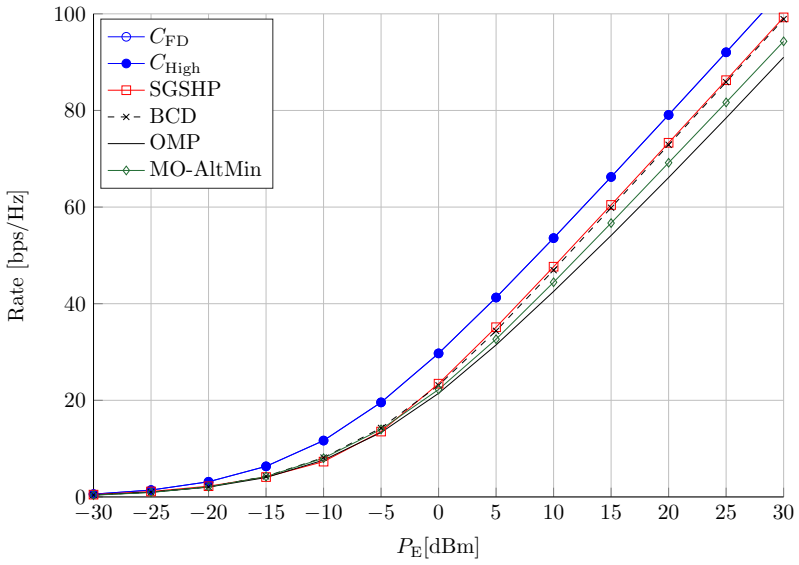

Figure 6.6.. The rates achievable with different precoding techniques with FI-HBF transmitter and receiver with 2 subarrays. The antenna setup is asymmetrical, with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx. Here, C_{FD} coincides with C_{High}.

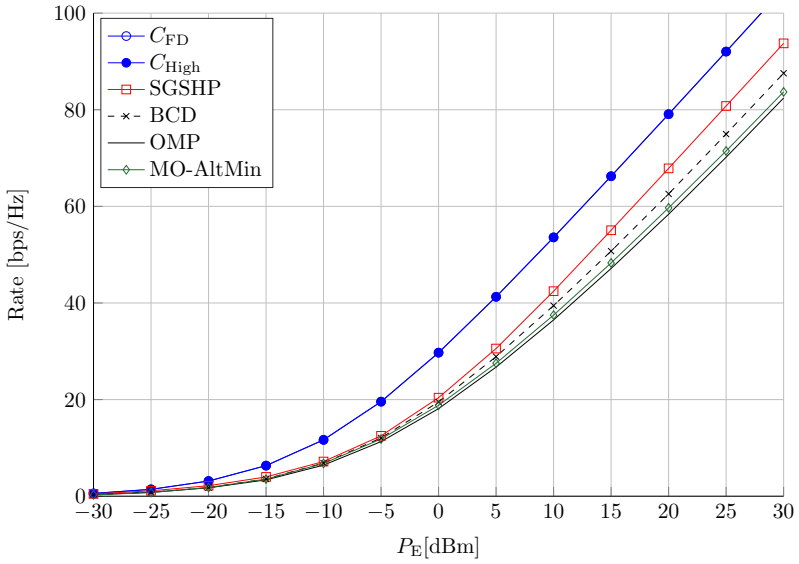

Figure 6.7.. The rates achievable with different precoding techniques with FI-HBF transmitter and receiver with 8 subarrays. The antenna setup is asymmetrical, with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx. Here, C_{FD} coincides with C_{High}.

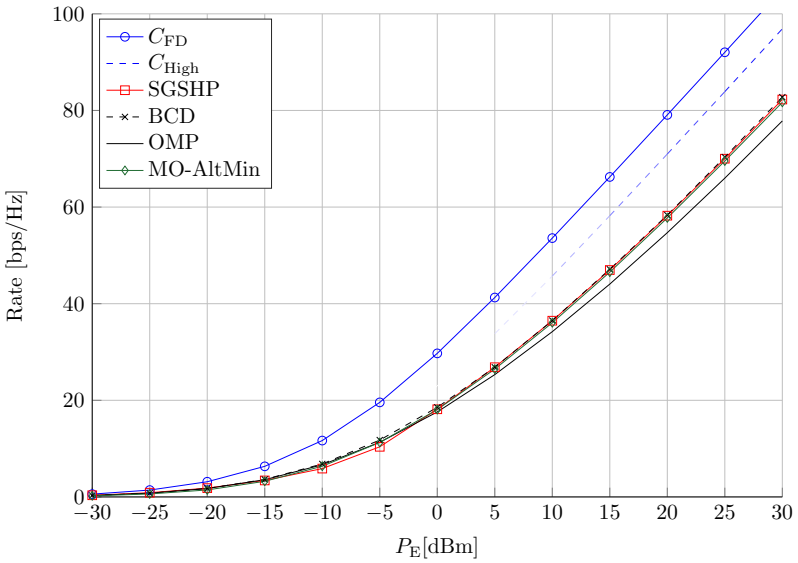

Figure 6.8.. The rates achievable with different precoding techniques with 4 subarrays at the transmitter and 4 at the receiver. The antenna setup is asymmetrical, with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx.

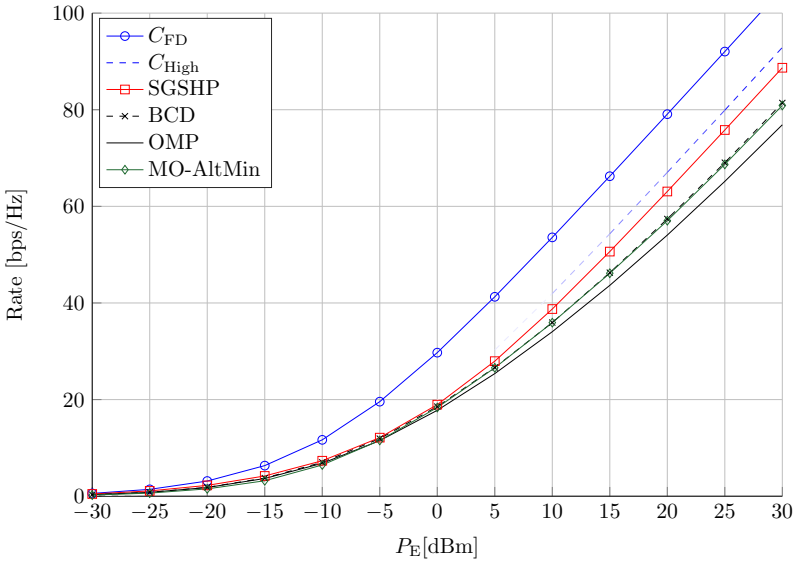

Figure 6.9.. The rates achievable with different precoding techniques with 8 subarrays at the transmitter and FI-HBF receiver. The antenna setup is asymmetrical, with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx.

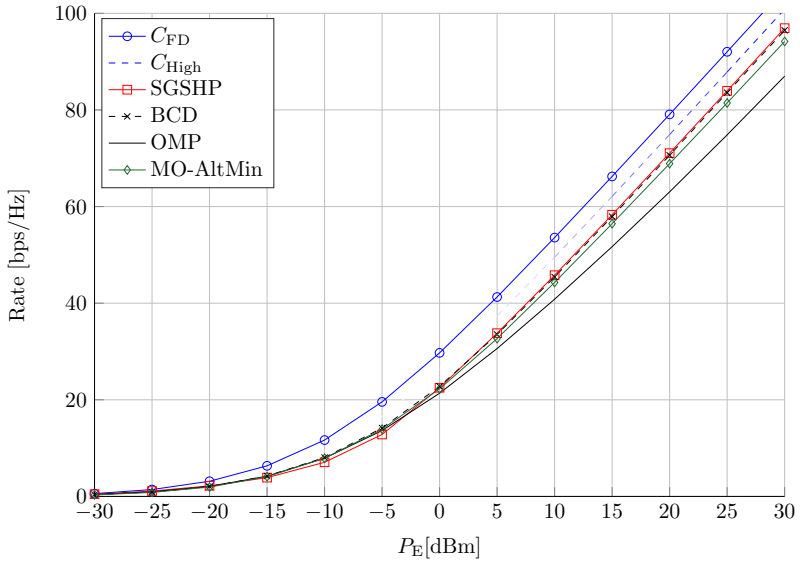

Figure 6.10.. The rates achievable with different precoding techniques with 2 subarrays at the transmitter and FI-HBF receiver. The antenna setup is asymmetrical, with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx.

We notice that the performance of the OMP algorithm is always inferior, which was reported, e.g., in [82]. The reason why SGSHP is sometimes better and sometimes equal to the SotA is not fully understood. It is also interesting, why very different algorithms happen to perform exactly the same.

Next, we take a closer look on the details of the SGSHP algorithms. In the description in Section 6.3.1 we pointed out that the algorithm may be tuned in order to match different system requirements or scenarios. More specifically, the tuning is achieved by designing the update of the projection matrices in each step of the algorithm—this, in turn, governs the channel subspace allocated to subsequent streams. Here, we show how the channel model may affect the choice. For this sake, we compare two proposals for the design of the projection matrices—the update of the projectors that assures no interference between the subsequent allocated streams (6.25) and the less conservative update (6.26). In the simulations, we consider two different channel models, namely the mmWave channel model which is described in Section 4.3 and the Rayleigh channel model. The Rayleigh channel model is the usual assumption for a non-line-of-sight channel at lower frequencies, a scenario that exhibits rich scattering and omnidirectivity. The results follow the expectation. For the spatially-sparse mmWave channel, the strict requirement regarding inter-stream

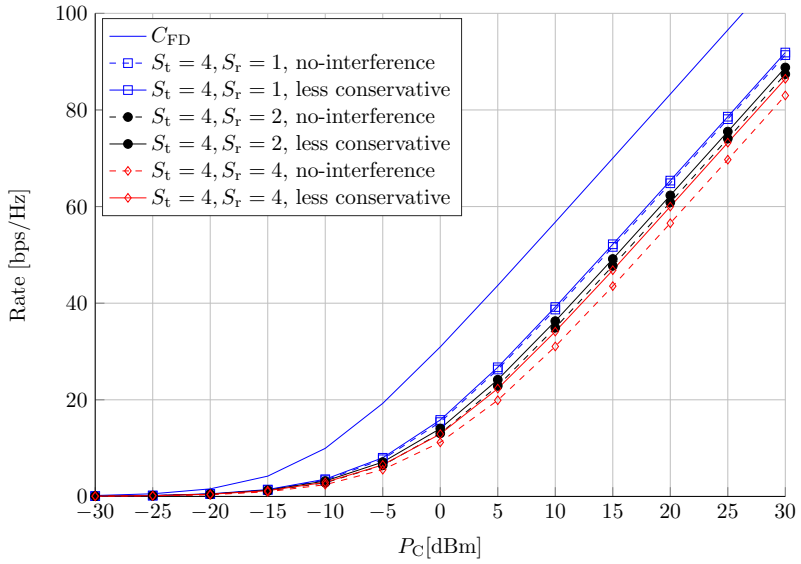

Figure 6.11.. Comparison between projection matrix update strategies for SGSHP and the Rayleigh channel. The "no-interference" label corresponds to the update like in (6.25) and "less conservative" to (6.26). The Tx is equipped with 128 antennas and 8 RF chains. The Rx is equipped with 32 antennas and 8 RF chains.

interference results in severe achievable rate drop, as presented in Fig. 6.12. This is attributed to the fact that the mmWave channel's number of significant singular values is low and the associated subspaces are occupied in the first steps of the algorithm. Consequently, the subsequent streams are assigned to non-relevant subspaces. The situation is different when the channel matrix is well conditioned (the singular values do not vary much). This is the case for the Rayleigh channel which models a rich scattering environment. There, the rates achieved with SGSHP using either of the updates do not differ much, as can be seen in Fig. 6.11. Nevertheless, the no-interference approach is still inferior in terms of the achievable rate. This behavior is not surprising taken into account that the achievable rate expression assumes optimal decoding at the receiver—a strategy that separates the streams in the receiver is not needed and will very likely negatively affect the performance. We note that this would change in the multiuser downlink scenario (not considered in this work), where the receivers do not cooperate during the decoding of the information.

In the end, we provide some results for low-SNR regime. We relate the performance of the aforementioned algorithms, including the novel low-SNR precoding scheme to the constrained capacity and novel low-SNR upper bound. The results in Figs. 6.13-

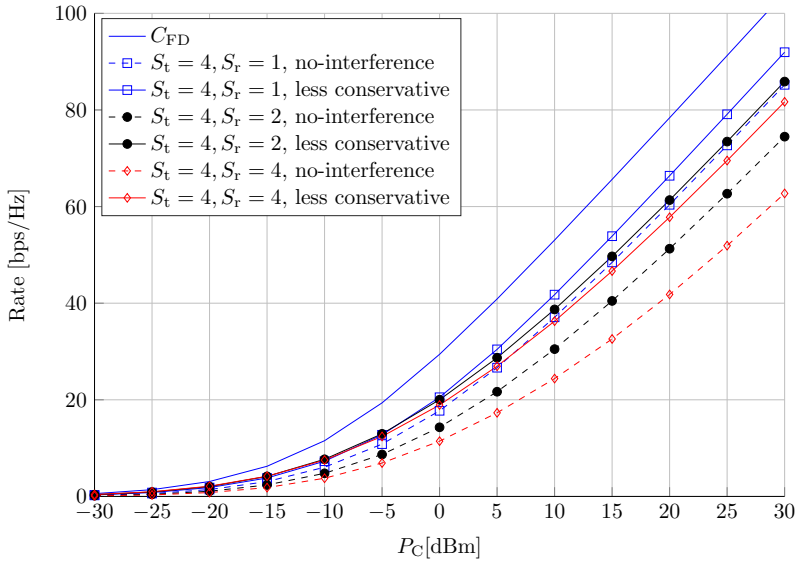

Figure 6.12.. Comparison between projection matrix update strategies for SGSHP and the mmWave channel. The "no-interference" label corresponds to the update like in (6.25) and "less conservative" to (6.26). The Tx is equipped with 128 antennas and 8 RF chains. The Rx is equipped with 32 antennas and 8 RF chains.

6.17 indicate superior operation of the low-SNR algorithm compared to the remaining SotA algorithms, with exclusion of the configuration where $S_{\mathrm{t}} = 1$, $S_{\mathrm{r}} = 2$. The SGSHP algorithm provides equal or slightly superior results in comparison to the SotA.

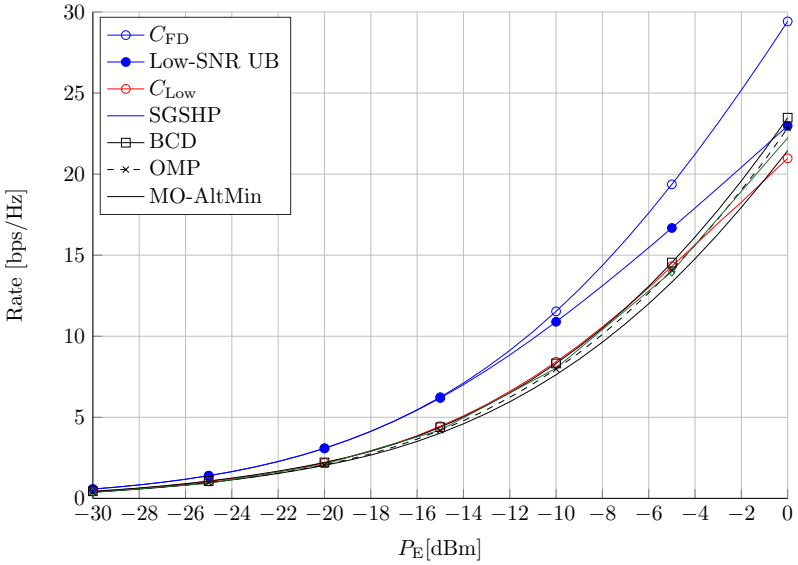

Figure 6.13.. The rates achievable at low SNR with different precoding techniques with FI-HBF transmitter and receiver with 2 subarrays. The antenna setup is asymmetrical, with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx.

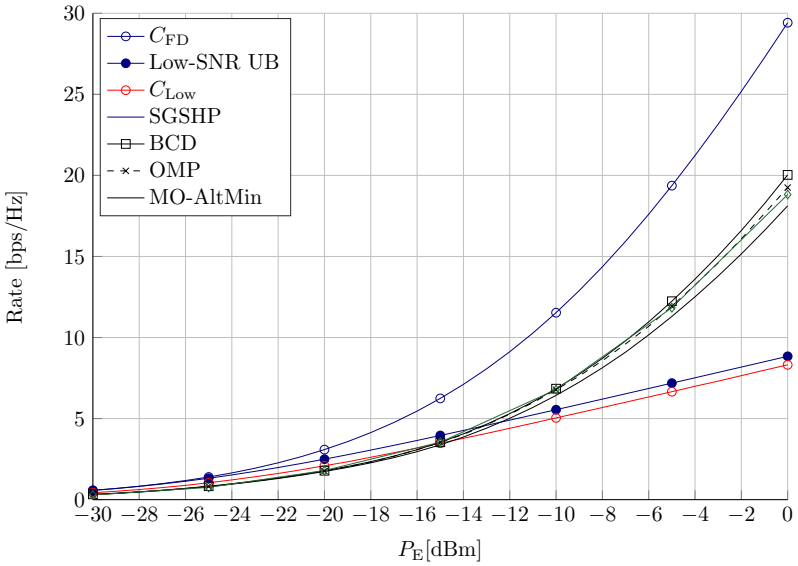

Figure 6.14.. The rates achievable at low SNR with different precoding techniques with FI-HBF transmitter and receiver with 8 subarrays. The antenna setup is asymmetrical, with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx.

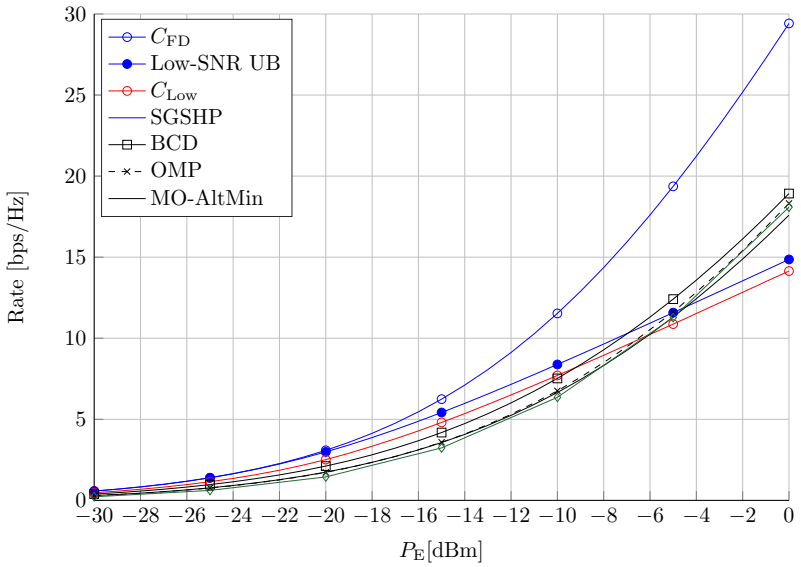

Figure 6.15.. The rates achievable at low SNR with different precoding techniques with 4 subarrays at the transmitter and 4 at the receiver. The antenna setup is asymmetrical, with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx.

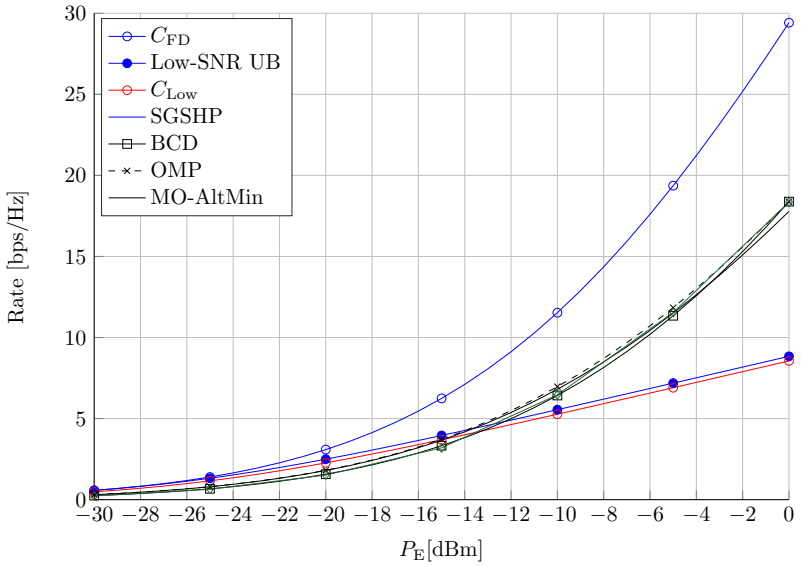

Figure 6.16.. The rates achievable at low SNR with different precoding techniques with 8 subarrays at the transmitter and FI-HBF receiver. The antenna setup is asymmetrical, with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx.

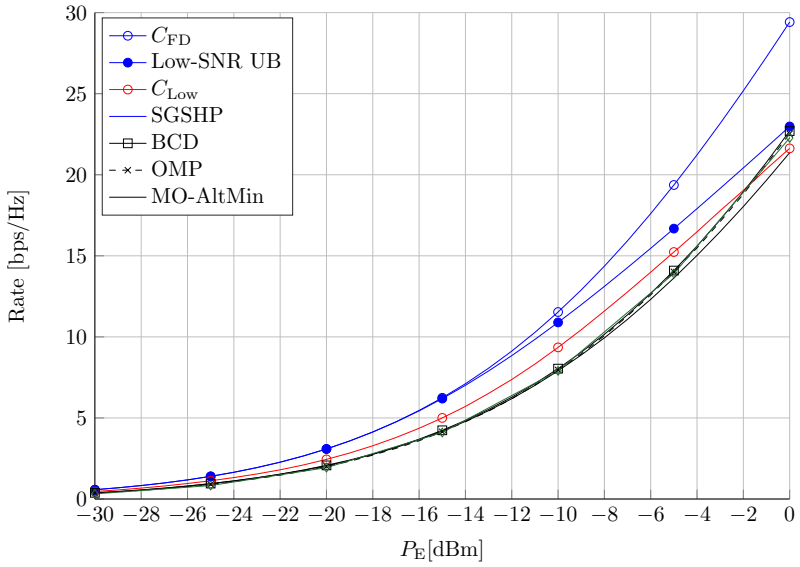

Figure 6.17.. The rates achievable at low SNR with different precoding techniques with 2 subarrays at the transmitter and FI-HBF receiver. The antenna setup is asymmetrical, with 128 Tx antennas, 32 Rx antennas, and 8 RF chains at the Tx and Rx.

Algorithm 4 Decomposition of the Optimal Fully Digital Precoder for SPI-HBF

Require: $\boldsymbol{P}_{\mathrm{FD}}^{\star}, \boldsymbol{\Gamma}, P_{\mathrm{TX}}$

Initialize $\left[\boldsymbol{P}_A^{(k),\star,0}\right]_{k=1}^{S_\mathrm{t}}, \left[\boldsymbol{P}_D^{(k),\star,0}\right]_{k=1}^{S_\mathrm{t}}$

repeat

  Update $\boldsymbol{P}_A^{(k)}$ independently for each k using any FI-HBF algorithm.

  Update $\boldsymbol{P}_D^{(k)}$:

  if Unconstrained update as in, e.g., [12, 22, 34] ($\boldsymbol{\Gamma} = \boldsymbol{0}$) **then**

    $\boldsymbol{P}_D^{(k)} = \boldsymbol{P}_A^{(k)^\dagger} \boldsymbol{P}_{\mathrm{FD}}^{\star,(k)} \; \forall k$

  else if emitted power constraint ($\Gamma = \boldsymbol{K}_\mathrm{A}, \; P_{\mathrm{TX}} = P_\mathrm{E}$) **then**

    $\boldsymbol{P}_D^{(k)} = \boldsymbol{P}_A^{(k)^\dagger} \boldsymbol{P}_{\mathrm{FD}}^{\star,(k)} \; \forall k$

    $\alpha = \sqrt{\frac{P_\mathrm{E}}{\sum_{k=1}^{S_\mathrm{t}} \|\boldsymbol{P}_A^{(k)} \boldsymbol{P}_D^{(k)}\|_\mathrm{F}^2}}$

    $\boldsymbol{P}_D^{(k)} = \alpha \boldsymbol{P}_D^{(k)} \; \forall k$

  else if consumed power constraint ($\Gamma = \boldsymbol{I}, \; P_{\mathrm{TX}} = P_\mathrm{C}$) **then**

    $\boldsymbol{P}_A = \mathrm{diag}(\left[\boldsymbol{P}_A^{(k)}\right]_{k=1}^{S_\mathrm{t}})$

    Update $\boldsymbol{P}_D$ like in (6.8)

    Decompose $\boldsymbol{P}_D$ into $\boldsymbol{P}_D^{(1)}, \ldots, \boldsymbol{P}_D^{(S_\mathrm{t})}$

  end if

until convergence

if Unconstrained update as in, e.g., [12, 22, 34] **then**

  $\alpha = \sqrt{\frac{P_\mathrm{E}}{\sum_{k=1}^{S_\mathrm{t}} \|\boldsymbol{P}_A^{(k)} \boldsymbol{P}_D^{(k)}\|_\mathrm{F}^2}}$

  $\boldsymbol{P}_D^{(k)} = \alpha \boldsymbol{P}_D^{(k)} \; \forall k$

end if

return $\boldsymbol{P}_A^{(k)}, \boldsymbol{P}_D^{(k)} \quad \forall k \in \{1, \ldots, S_\mathrm{t}\}$

Algorithm 5 SGSHP algorithm

Inputs:
$\boldsymbol{H}_{l,k}\ \forall k,l$
Initialize:
$\forall l \in \{1,\dots,S_{\mathrm{r}}\}, k \in \{1,\dots,S_{\mathrm{t}}\}$
$\boldsymbol{G}_A^{(l)} \leftarrow$ Null, $\boldsymbol{P}_A^{(k)} \leftarrow$ Null
$\boldsymbol{\Xi}_k^{(1)} \leftarrow \boldsymbol{I}$, $\boldsymbol{\Pi}_l^{(1)} \leftarrow \boldsymbol{I} \quad \forall k,l$
for $i = 1,\dots,N_{\mathrm{RF}}$ **do**
  if $i == 1$ **then**
    Get $\{k_1, l_1, \boldsymbol{p}_1, \boldsymbol{g}_1\}$ from solving (6.16)
  else
    Get $\{k_i, l_i, \boldsymbol{p}_i, \boldsymbol{g}_i\}$ from solving (6.23)
  end if
  $\boldsymbol{G}_A^{(l_i)} = \left[\boldsymbol{G}_A^{(l_i)} \boldsymbol{g}_i\right], \quad \boldsymbol{P}_A^{(k_i)} = \left[\boldsymbol{P}_A^{(k_i)} \boldsymbol{p}_i\right]$
  Update $\boldsymbol{\Xi}_k^{(i)}$, $\boldsymbol{\Pi}_l^{(i)} \quad \forall k,l$
end for
$\boldsymbol{G}_A = \mathrm{diag}\left(\boldsymbol{G}_A^{(1)},\dots,\boldsymbol{G}_A^{(\mathrm{S_r})}\right)$
$\boldsymbol{P}_A = \mathrm{diag}\left(\boldsymbol{P}_A^{(1)},\dots,\boldsymbol{P}_A^{(\mathrm{S_t})}\right)$ **return** $\boldsymbol{P}_A, \boldsymbol{G}_A$

Algorithm 6 Low-SNR precoding

Require: $\boldsymbol{H},\ N_{\mathrm{t}}^{\mathrm{RF},k}\ \forall k$

$N'_{\mathrm{RF}} \leftarrow \min_k\{N_{\mathrm{t}}^{\mathrm{RF},k}\}$
Determine $\boldsymbol{P}'_A$ and $\boldsymbol{P}'_D$ for FI-HBF with N'_{RF} RF chains. Utilize a well-established algorithm (e.g., BCD, OMP, MO-AltMin)
Determine $\boldsymbol{P}''_A$ and $\boldsymbol{P}''_D$ as presented in Theorem 1.
$\boldsymbol{P}_A \leftarrow \boldsymbol{P}''_A,\ \boldsymbol{P}_D \leftarrow \boldsymbol{P}''_D$
return $\boldsymbol{P}_A, \boldsymbol{P}_D$

7

Summary

In this work, we performed a thorough analysis of the subarray hybrid architectures. First, we derived new upper bounds for the achievable rates. Using them, we were able to show that in certain setups the available precoding techniques perform close to the limit.

Furthermore, we proposed new precoding techniques: the SGSHP and low-SNR precoding. SGSHP was shown to outperform the SotA algorithms in many subarray setups and to perform comparably in others. The low-SNR precoding has been shown to outperform the SotA algorithms in any investigated setup. We also showed under which conditions capacity is achievable with subarray hybrid architecture.

Moreover, we addressed the issue of updating the digital precoding matrix, if an alternating optimization of analog and digital matrices is employed. We showed that in many cases the scaling of the unconstrained solution with a constant (perceived as suboptimal) is actually optimal.

Finally, we advocate in our work for fair comparisons between the FI-HBF and SPI-HBF structures. For this sake, we proposed to use the consumed power as the power constraint, instead of the emitted power. This resulted in an advantage for the SPI-HBF structures, as less power is dissipated in their analog stages. While this may be not the only (or the most accurate) way of introducing fairness into comparing very complex FI-HBF structures to simpler SPI-HBF transceivers, it stresses the necessity of taking the issue into account, which is often neglected.

In our work we did not consider many important topics in mmWave precoding—e.g., the difficulties in obtaining the full channel information or imperfections of the

analog stage. Instead, we concentrated on discussing the limits of the SPI-HBF architectures and showing that with a fair comparison, there is no performance loss with respect to the FI-HBF architectures.

A

MIMO Capacity with Reduced Number of Streams

Theorem 3. The capacity C_K of a Gaussian MIMO point-to-point channel with number of streams not exceeding K, M transmit and N receive antennas, average power not exceeding P and known effective channel matrix $\boldsymbol{H'} = \boldsymbol{R}_\eta^{-1/2}\boldsymbol{H}$, where $\boldsymbol{R}_\eta$ is the noise covariance matrix, reads

$$\log_2 \det\left(\boldsymbol{I} + \boldsymbol{H'}^{\mathrm{H}}\boldsymbol{H'}\boldsymbol{R}^\star\right) \tag{A.1}$$

with $\boldsymbol{R}^\star = \boldsymbol{U}_{:,1:K}\boldsymbol{L}\boldsymbol{U}_{:,1:K}^{\mathrm{H}}$, where $\boldsymbol{U}_{:,1:K}$ is a matrix consisting of K eigenvectors of $\boldsymbol{H'}^{\mathrm{H}}\boldsymbol{H'}$ corresponding to the K largest eigenvalues of the matrix, and $\boldsymbol{L}$ is a diagonal power allocation matrix obtained through the waterfilling algorithm on the K largest eigenvalues of $\boldsymbol{H'}^{\mathrm{H}}\boldsymbol{H'}$.

Proof. The channel's capacity can be obtained through solving an optimization problem that reads

$$\begin{aligned} \boldsymbol{C}_K = \max_{\boldsymbol{R}} & \log_2 \det\left(\boldsymbol{I} + \boldsymbol{H'}^{\mathrm{H}}\boldsymbol{H'}\boldsymbol{R}\right) \\ & \text{s.t. } \operatorname{tr}(\boldsymbol{R}) \leq P,\ \operatorname{rank}(\boldsymbol{R}) \leq K \\ & \boldsymbol{R} \succeq 0. \end{aligned} \tag{A.2}$$

Further, we write the eigenvalue decomposition of $\boldsymbol{H'}^{\mathrm{H}}\boldsymbol{H'}$ and $\boldsymbol{R}$ as $\boldsymbol{H'}^{\mathrm{H}}\boldsymbol{H'} = \boldsymbol{U}\boldsymbol{\Xi}\boldsymbol{U}^{\mathrm{H}}$ and $\boldsymbol{R} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{H}}$, respectively. The diagonal entries of $\boldsymbol{\Xi}$ and $\boldsymbol{\Lambda}$ are arranged in non-increasing order, i.e., $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_M$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M$.

In order to account for the rank constraint, only K of the diagonal entries of $\boldsymbol{\Lambda}$ are nonzero. Consequently, we write (A.2) equivalently as

$$\begin{aligned} \boldsymbol{C}_K = \max_{\boldsymbol{U},\boldsymbol{\Lambda}} \log_2 \det\left(\boldsymbol{I} + \boldsymbol{U}\boldsymbol{\Xi}\boldsymbol{U}^{\mathrm{H}}\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{H}}\right) \\ \text{s.t. } \operatorname{tr}(\boldsymbol{\Lambda}) \leq P,\ \|\operatorname{diag}(\boldsymbol{\Lambda})\|_0 \leq K \end{aligned} \tag{A.3}$$

From Hadamard inequality we know that if $\boldsymbol{A}$ is positive semidefinite then $\det(\boldsymbol{A}) \leq \prod \boldsymbol{A}_{i,i}$ with equality if $\boldsymbol{A}$ is diagonal. Therefore we get $\boldsymbol{V}^\star = \boldsymbol{U}$ and (A.2) is equivalent to a convenient scalar form:

$$\boldsymbol{C}_K = \max_{\{\lambda\}} \sum_{i=1}^{\min\{\operatorname{rank}(\boldsymbol{H}'),K\}} \log(1+\xi_i\lambda_i) \tag{A.4}$$

Such problem can be solved by means of the waterfilling algorithm [24, 71, 74] and the solution reads

$$\lambda_i = \left(\mu - \frac{1}{\xi_i}\right)^+ \tag{A.5}$$

where $x^+ = \max(x, 0)$ and μ is such that $\sum_i \lambda_i = P$. Finally, we write the optimal signal covariance matrix as $\boldsymbol{R}^\star = \boldsymbol{U}\boldsymbol{\Lambda}^\star\boldsymbol{U}^{\mathrm{H}}$. We know from the design parameters that $\lambda_{K+1} = \cdots = \lambda_M = 0$ and therefore we can reduce the expression to $\boldsymbol{R}^\star = \boldsymbol{U}_{:,1:K}\boldsymbol{\Lambda}_{1:K,1:K}\boldsymbol{U}^{\mathrm{H}}_{:,1:K} = \boldsymbol{U}_{:,1:K}\boldsymbol{L}\boldsymbol{U}^{\mathrm{H}}_{:,1:K}$. □

We note that the theorem is a straightforward corollary of the MIMO capacity derivation from [71] and the proof follows similar arguments.

Bibliography

[1] http://www.anokiwave.com/. Accessed: 2018-04-18.

[2] IEEE Std 802.11™-2016. Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications.

[3] WirelessHD Specification Overview. Tech. Rep., 2010.

[4] A. Adhikary, E. Al Safadi, M. Samimi, R. Wang, G. Caire, T. Rappaport, and A. Molisch. Joint Spatial Division and Multiplexing for mm-Wave Channels. *IEEE J. Sel. Areas Commun.*, 32(6):1239–1255, June 2014.

[5] M. Akdeniz, Y. Liu, M. Samimi, S. Sun, S. Rangan, T. Rappaport, and E. Erkip. Millimeter Wave Channel Modeling and Cellular Capacity Evaluation. *IEEE J. Sel. Areas Commun.*, 32(6):1164—1179, June 2014.

[6] S. M. Alamouti. A simple transmit diversity technique for wireless communications. *IEEE J. Sel. Areas Commun.*, 16(8):1451–1458, Oct 1998.

[7] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath. Channel Estimation and Hybrid Precoding for Millimeter Wave Cellular Systems. *IEEE J. Sel. Topics Signal Process.*, 8(5):831–846, Oct. 2014.

[8] A. Alkhateeb, R. W. Heath, and G. Leus. Achievable rates of multi-user millimeter wave systems with hybrid precoding. In *2015 IEEE International Conference on Communication Workshop (ICCW)*, pages 1232–1237, June 2015.

[9] A. Alkhateeb and R. W. Heath Jr. Frequency selective hybrid precoding for limited feedback millimeter wave systems. *CoRR*, abs/1510.00609, 2015.

[10] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. E. Lozano, A. C. K. Soong, and J. C. Zhang. What will 5g be? *CoRR*, abs/1405.2957, 2014.

[11] O. E. Ayach, R. W. Heath, S. Rajagopal, and Z. Pi. Multimode precoding in millimeter wave MIMO transmitters with multiple antenna sub-arrays. In o, editor, *Proc. IEEE Global Communications Conf. (GLOBECOM)*, pages 3476–3480, Dec. 2013. subarrays.

[12] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath. Spatially sparse precoding in millimeter wave MIMO systems. *IEEE Trans. Wireless Commun.*, 13(3):1499–1513, Mar. 2014.

[13] C. N. Barati, S. A. Hosseini, M. Mezzavilla, T. Korakis, S. S. Panwar, S. Rangan, and M. Zorzi. Initial access in millimeter wave cellular systems. *IEEE Trans. Wireless Commun.*, 15(12):7926–7940, Dec 2016.

[14] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and H. V. Poor. *MIMO Wireless Communications*. Cambridge University Press, New York, NY, USA, 2007.

[15] S. Blandino, G. Mangraviti, C. Desset, A. Bourdoux, P. Wambacq, and S. Pollin. Multi-user hybrid mimo at 60 ghz using 16-antenna transmitters. *IEEE J. Sel. Topics Signal Process.*, 66(2):848–858, Feb 2019.

[16] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski. Five disruptive technology directions for 5g. *IEEE Commun. Mag.*, 52(2):74–80, February 2014.

[17] G. Foschini and M. Gans. On limits of wireless communications in a fading environment when using multiple antennas. *Wireless Personal Communications*, 6(3):311–335, Mar 1998.

[18] X. Gao, L. Dai, S. Han, C. L. I, and R. W. Heath. Energy-efficient hybrid analog and digital precoding for mmwave MIMO systems with large antenna arrays. *IEEE J. Sel. Areas Commun.*, 34(4):998–1009, Apr. 2016.

[19] X. Gao, L. Dai, S. Han, C. L. I, and R. W. Heath. Energy-efficient hybrid analog and digital precoding for mmwave MIMO systems with large antenna arrays. *IEEE J. Sel. Areas Commun.*, 34(4):998–1009, Apr. 2016.

[20] Y. Gao, M. Khaliel, F. Zheng, and T. Kaiser. Rotman lens based hybrid analog –digital beamforming in massive MIMO systems: Array architectures, beam

selection algorithms and experiments. *IEEE Trans. Veh. Technol.*, 66(10):9134–9148, Oct. 2017.

[21] A. Garcia-Rodriguez, V. Venkateswaran, P. Rulikowski, and C. Masouros. Hybrid analog-digital precoding revisited under realistic rf modeling. *IEEE Wireless Commun. Lett.*, 5(5):528–531, Oct 2016.

[22] H. Ghauch, T. Kim, M. Bengtsson, and M. Skoglund. Subspace estimation and decomposition for large millimeter-wave MIMO systems. *IEEE J. Sel. Topics Signal Process.*, 10(3):528–542, Apr. 2016.

[23] M. Giordani, M. Mezzavilla, and M. Zorzi. Initial access in 5g mmwave cellular networks. *IEEE Commun. Mag.*, 54(11):40–47, November 2016.

[24] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath. Capacity limits of MIMO channels. *IEEE J. Sel. Areas Commun.*, 21(5):684–702, June 2003.

[25] A. Grebennikov. Power combiners, impedance transformers and directional couplers. In *High Freq. Electron.*, volume 6, Dec. 2007.

[26] C. Guthy, W. Utschick, and G. Dietl. Low-Complexity Linear Zero-Forcing for the MIMO Broadcast Channel. *IEEE J. Sel. Topics Signal Process.*, 3(6):1106–1117, Dec. 2009.

[27] S. Hafner, D. Dupleich, R. Muller, J. Luo, E. Schulz, C. Schneider, R. Thoma, X. Lu, and T. Wang. Characterisation of Channel Measurements at 70GHz in Indoor Femtocells. In *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, pages 1—5, May 2015.

[28] S. Han, C. l. I, Z. Xu, and C. Rowell. Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G. *IEEE J. Sel. Areas Commun.*, 53(1):186–194, Jan. 2015. subarrays.

[29] S. He, C. Qi, Y. Wu, and Y. Huang. Energy-efficient transceiver design for hybrid sub-array architecture MIMO systems. *IEEE Access*, 4:9895–9905, 2016.

[30] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed. An overview of signal processing techniques for millimeter wave MIMO systems. *IEEE J. Sel. Topics Signal Process.*, 10(3):436–453, Apr. 2016.

[31] X. Huang, Y. J. Guo, and J. D. Bunton. A hybrid adaptive antenna array. *IEEE Trans. Wireless Commun.*, 9(5):1770–1779, May 2010.

[32] R. Hunger, D. A. Schmidt, and W. Utschick. Sum-capacity and MMSE for the MIMO broadcast channel without eigenvalue decompositions. In *2007 IEEE International Symposium on Information Theory*, pages 776–780, June 2007.

[33] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh. Millimeter wave beamforming for wireless backhaul and access in small cell networks. *IEEE Trans. Commun.*, 61(10):4391–4403, October 2013.

[34] S. S. Ioushua and Y. C. Eldar. Hybrid analog-digital beamforming for massive MIMO systems. *CoRR*, abs/1712.03485, 2017.

[35] M. Iwanow, S. Bazzi, and W. Utschick. A greedy approach for multiuser mmwave hybrid precoding with subarray architectures. In *22nd International ITG Workshop on Smart Antennas (WSA 2018)*, Bochum, Germany, Mar. 2018.

[36] M. Iwanow, N. Vucic, S. Bazzi, J. Luo, and W. Utschick. A greedy approach for mmwave hybrid precoding with subarray architectures. In *2017 51th Asilomar Conference on Signals, Systems and Computers*, Nov. 2017.

[37] M. Iwanow, N. Vucic, M. Castaneda, J. Luo, W. Xu, and W. Utschick. Some aspects on hybrid wideband transceiver design for mmWave communication systems. In *Proc. WSA 2016; 20th Int. ITG Workshop Smart Antennas*, pages 1–8, Munich, Mar. 2016.

[38] M. Iwanow, N. Vucic, W. Utschick, M. Castañeda, W. Xu, and J. Luo. Data rate bound for mmWave hybrid beamforming systems with subarrays. In *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC) (IEEE SPAWC 2017)*, Sapporo, Japan, July 2017.

[39] M. Joham, W. Utschick, and J. A. Nossek. Linear transmit processing in MIMO communications systems. *IEEE Trans. Signal Process.*, 53(8):2700–2712, Aug. 2005.

[40] N. Kwak. Principal component analysis based on l1-norm maximization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(9):1672–1680, Sept 2008.

[41] S. Lee, A. P. Chandrakasan, and H. S. Lee. A 1 gs/s 10b 18.9 mw time-interleaved sar adc with background timing skew calibration. *IEEE J. Solid-State Circuits*, 49(12):2846–2856, Dec 2014.

[42] Y. Li, E. Pateromichelakis, N. Vucic, J. Luo, W. Xu, and G. Caire. Radio resource management considerations for 5g millimeter wave backhaul and access networks. *IEEE Commun. Mag.*, 55(6):86–92, June 2017.

[43] A. Liu and V. Lau. Phase Only RF Precoding for Massive MIMO Systems With Limited RF Chains. *IEEE Trans. Signal Process.*, 62(17):4505–4515, Sept. 2014.

[44] R. Mendez-Rial, C. Rusu, A. Alkhateeb, N. Gonzalez-Prelcic, and R. Heath. Channel estimation and hybrid combining for mmWave: Phase shifters or switches? In *Information Theory and Applications Workshop (ITA), 2015*, pages 90–97, Feb. 2015.

[45] A. Mezghani and J. A. Nossek. On ultra-wideband mimo systems with 1-bit quantized outputs: Performance analysis and input optimization. In *2007 IEEE International Symposium on Information Theory*, pages 1286–1289, June 2007.

[46] R. Méndez-Rial, C. Rusu, N. González-Prelcic, A. Alkhateeb, and R. W. Heath. Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches? *IEEE Access*, 4:247–267, 2016.

[47] J. Mo, A. Alkhateeb, S. Abu-Surra, and R. W. Heath. Hybrid architectures with few-bit adc receivers: Achievable rates and energy-rate tradeoffs. *IEEE Trans. Wireless Commun.*, 16(4):2274–2287, April 2017.

[48] J. Mo and R. W. Heath. Capacity analysis of one-bit quantized MIMO systems with transmitter channel state information. *IEEE Trans. Signal Process.*, 63(20):5498–5512, Oct. 2015.

[49] A. F. Molisch, A. Karttunen, R. Wang, C. U. Bas, S. Hur, J. Park, and J. Zhang. Millimeter-wave channels in urban environments. In *2016 10th European Conference on Antennas and Propagation (EuCAP)*, pages 1–5, April 2016.

[50] 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. *CoRR*, abs/1609.05078, 2016.

[51] W. Ni, X. Dong, and W. Lu. Near-optimal hybrid processing for massive mimo systems via matrix decomposition. *IEEE Trans. Signal Process.*, 65(15):3922–3933, Aug 2017.

[52] D. P. Palomar. *A Unified Framework for Communications through MIMO Channels*. Ph. D. Dissertationa, Universitat Politecnica de Catalunya, Barcelona, May 2003.

[53] S. Park, A. Alkhateeb, and R. W. Heath. Dynamic subarrays for hybrid precoding in wideband mmwave mimo systems. *IEEE Trans. Wireless Commun.*, 16(5):2907–2920, May 2017.

[54] S. Park, A. Alkhateeb, and R. W. Heath Jr. Dynamic Subarrays for Hybrid Precoding in Wideband mmWave MIMO Systems. *arXiv:1606.08405 [cs, math]*, June 2016. arXiv: 1606.08405.

[55] S. Payami, M. Ghoraishi, and M. Dianati. Hybrid beamforming with reduced number of phase shifters for massive MIMO systems. *CoRR*, abs/1706.01304, 2017.

[56] D. M. Pozar. *Microwave engineering*. John Wiley & Sons, 2009.

[57] T. Rappaport, R. Heath, R. Daniels, and J. Murdock. *Millimeter wave wireless communications*. Prentice Hall, 2015. Includes bibliographical references (pages 585-651) and index.

[58] T. S. Rappaport, G. R. MacCartney, M. K. Samimi, and S. Sun. Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design. *IEEE Trans. Commun.*, 63(9):3029–3056, Sep. 2015.

[59] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez. Millimeter Wave Mobile Communications for 5g Cellular: It Will Work! *IEEE Access*, 1:335–349, 2013.

[60] K. Roth and J. A. Nossek. Achievable rate and energy efficiency of hybrid and digital beamforming receivers with low resolution adc. *IEEE J. Sel. Areas Commun.*, 35(9):2056–2068, Sep. 2017.

[61] A. Saleh and R. Valenzuela. A Statistical Model for Indoor Multipath Propagation. *IEEE J. Sel. Areas Commun.*, 5(2):128—137, Feb. 1987.

[62] M. Samimi and T. Rappaport. Ultra-wideband statistical channel model for non line of sight millimeter-wave urban channels. In *2014 IEEE Global Communications Conference (GLOBECOM)*, pages 3483—3489, Dec. 2014.

[63] M. K. Samimi and T. S. Rappaport. 3-D Statistical Channel Model for Millimeter-Wave Outdoor Mobile Broadband Communications. *arXiv preprint arXiv:1503.05619*, 2015.

[64] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–, july, october 1948.

[65] J. Singh and S. Ramakrishna. On the feasibility of codebook-based beamforming in millimeter wave systems with multiple antenna arrays. *IEEE Trans. Wireless Commun.*, 14(5):2670–2683, May 2015.

[66] F. Sohrabi and W. Yu. Hybrid analog and digital beamforming for OFDM-based large-scale MIMO systems. In *Proc. IEEE 17th Int. Workshop Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–6, July 2016.

[67] F. Sohrabi and W. Yu. Hybrid digital and analog beamforming design for large-scale antenna arrays. *IEEE J. Sel. Topics Signal Process.*, 10(3):501–513, Apr. 2016.

[68] F. Sohrabi and W. Yu. Hybrid analog and digital beamforming for mmwave OFDM large-scale antenna arrays. *IEEE J. Sel. Areas Commun.*, 35(7):1432–1443, July 2017.

[69] N. Song, T. Yang, and H. Sun. Overlapped subarray based hybrid beamforming for millimeter wave multiuser massive mimo. *IEEE Signal Process. Lett.*, 24(5):550–554, May 2017.

[70] P. Sudarshan, N. B. Mehta, A. F. Molisch, and J. Zhang. Channel statistics-based rf pre-processing with antenna selection. *IEEE Trans. Wireless Commun.*, 5(12):3501–3511, December 2006.

[71] E. Telatar. Capacity of multi-antenna gaussian channels. *Eur. Trans. Telecommun.*, 10(6):585–595, 1999.

[72] T. Thomas, H. C. Nguyen, G. R. MacCartney, T. S. Rappaport, and others. 3D mmWave channel model proposal. In *Vehicular Technology Conference (VTC Fall), 2014 IEEE 80th*, pages 1–6. IEEE, 2014.

[73] Y. M. Tsang, A. S. Y. Poon, and S. Addepalli. Coding the beams: Improving beamforming training in mmwave communication system. In *2011 IEEE Global Telecommunications Conference - GLOBECOM 2011*, pages 1–6, Dec 2011.

[74] D. Tse and P. Viswanath. *Fundamentals of Wireless Communication*. Cambridge University Press, 2004.

[75] C. G. Tsinos, S. Maleki, S. Chatzinotas, and B. Ottersten. On the energy-efficiency of hybrid analog digital transceivers for single- and multi-carrier large antenna array systems. *IEEE J. Sel. Areas Commun.*, 35(9):1980–1995, Sept 2017.

[76] W. Utschick, C. Stockle, M. Joham, and J. Luo. Hybrid lisa precoding for multiuser millimeter-wave communications. *IEEE Trans. Wireless Commun.*, PP(99):1–1, 2017.

[77] 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.*, 58(8):4131–4143, Aug 2010.

[78] S. Vishwanath, N. Jindal, and A. Goldsmith. Duality, achievable rates, and sum-rate capacity of gaussian MIMO broadcast channels. *IEEE Trans. Inf. Theory*, 49(10):2658–2668, Oct. 2003.

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

[80] F. Willems and E. van der Meulen. The discrete memoryless multiple-access channel with cribbing encoders. *IEEE Trans. Inf. Theory*, 31(3):313–327, May 1985.

[81] Z. Xu, S. Han, Z. Pan, and C. L. I. Alternating beamforming methods for hybrid analog and digital MIMO transmission. In *Proc. IEEE Int. Conf. Communications (ICC)*, pages 1595–1600, June 2015.

[82] X. Yu, J. C. Shen, J. Zhang, and K. B. Letaief. Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems. *IEEE J. Sel. Topics Signal Process.*, 10(3):485–500, Apr. 2016.

[83] Y. Zeng and R. Zhang. Millimeter wave mimo with lens antenna array: A new path division multiplexing paradigm. *IEEE Trans. Commun.*, 64(4):1557–1571, April 2016.

[84] Y. Zeng and R. Zhang. Cost-effective millimeter-wave communications with lens antenna array. *IEEE Wireless Commun.*, 24(4):81–87, Aug. 2017.

[85] E. Zhang and C. Huang. On achieving optimal rate of digital precoder by rf-baseband codesign for mimo systems. In *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, pages 1–5, Sept 2014.

[86] X. Zhang, A. F. Molisch, and S.-Y. Kungand. Variable-phase-shift-based rf-baseband codesign for mimo antenna selection. *IEEE Trans. Signal Process.*, 53(11):4091–4103, Nov 2005.

[87] L. Zheng and D. N. C. Tse. Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels. *IEEE Trans. Inf. Theory*, 49(5):1073–1096, May 2003.

[88] L. Zhou and Y. Ohashi. Efficient codebook-based mimo beamforming for millimeter-wave wlans. In *2012 IEEE 23rd International Symposium on Personal, Indoor and Mobile Radio Communications - (PIMRC)*, pages 1885–1889, Sep. 2012.

[89] F. Zhu, S. He, R. Li, Y. Huang, and X. You. Energy-efficient hybrid precoding for broadband millimeter wave communication systems. In *Proc. 9th Int. Conf. Wireless Communications and Signal Processing (WCSP)*, pages 1–5, Oct. 2017.