TECHNISCHE UNIVERSITÄT MÜNCHEN

Fachgebiet Methoden der Signalverarbeitung

Principles and Algorithms for Transmission in Multiple-Input Multiple-Output Broadband Multiuser Systems

Pedro Tejera Palomares

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: Univ.-Prof. R. Kötter, Ph.D.

Prüfer der Dissertation:

1. Unv.-Prof. Dr. W. Utschick
2. Prof. Dr. H. Bölcskei,
   Eidgenössische Technische Hochschule Zürich/Schweiz

Die Dissertation wurde am 27.05.2008 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 24.07.2008 angenommen.

Acknowledgements

I would like to thank everybody I had the opportunity to meet, work with or simply spend some time with during my time at the Technische Universität München (TUM), since, in the end, in one way or the other, they all left an imprint in this work. Of course, the same can be said of all my relatives, friends and acquaintances in the private sphere. To them I am most grateful.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Scope and contributions

Increasing demand for broadband services calls for higher data rates in future wireless communication systems [109]. Data rates of up to 100 Mb/s for high mobility and wide area coverage and up to 1 Gb/s for low mobility and local area coverage are expected in fourth generation systems [138, 3]. In the way to such transmission rates there are two major barriers to be overcome. The first is the scarcity of spectrum, which limits the amount of bandwidth available for transmission. The second is the wireless channel that severely distorts the signal due to multipath propagation. The combination of multiple antennas and multicarrier technology seems key in enabling achievability of the expected rates under the mentioned constraints [98, 138, 44]. On the one hand, multiple-input multiple-output (MIMO) channels resulting from the use of multiple antennas at both transmitter and receiver show higher capacity than single-input single-output (SISO) channels and, at high signal-to-noise ratios (SNR), this difference linearly grows with the rank of the MIMO channel matrix. Thus, multiple antennas lead to higher spectral efficiency. On the other hand, multicarrier techniques, such as orthogonal frequency-division multiplexing (OFDM), transform the frequency selective broadband channel into a set of nearly flat narrowband channels. As a result, distortion due to multipath is reduced and equalization at the receiver is greatly simplified. These technologies have already been embraced in ongoing standardarization activities for future wireless systems. These include fixed wireless access networks for the last mile, wireless local area networks and cellular mobile networks. Also in modern digital subscriber line communication systems plays the combination of MIMO and multicarrier technologies a key role [26, 27]. In these systems, the use of high frequencies causes significant electromagnetic coupling between neighboring twisted-pairs within a binder group, which is commonly known as crosstalk and gives rise to an effective MIMO channel. In addition, high frequencies in transmission also causes the channel to exhibit severe frequency selectivity. This motivates the use of multicarrier technology in order to keep equalization simple and adapt to the selective spectral characteristic of the channel through adequate bit- and power-loading schemes.

The focus of this work is on transmission schemes in point-to-multipoint MIMO-OFDM communication systems. That is, communication systems are considered in which a transmitter sends information to a number of receivers or users. These systems are in the information-theoretic literature commonly known as broadcast channels. Information sent to each user is independent of the information sent to any other user and users can not cooperate with each other in order to perform detection. The transmitter transmits information over multiple inputs and each receiver receives information over multiple outputs. Inputs and outputs will be generally referred to as antennas although application of the principles and algorithms presented and discussed in this work are generally also applicable

to wired communication systems. The underlying transmission scheme is assumed to be OFDM. All along this work we shall assume that the transmitter has perfect knowledge of the channel matrices of all users in the system and each receiver perfectly knows its own channel matrix. Correspondingly, the focus will be on algorithms that exploit this knowledge available at the transmitter in order to adapt to the channel rather than on algorithms that leverage diversity in order to bridge uncertainty about the channel state.

Chapter 1 starts reviewing fundamental results on general broadcast channels, i.e., broadcast channels defined by generic alphabets and probability transition functions. This part has a tutorial character and includes classical results from the late seventies and early eighties such as the Marton achievability region. In the second part of this chapter we turn our attention to recent results concerning MIMO Gaussian broadcast channels, i.e., broadcast channels with multiple antennas and Gaussian probability transition functions. Relating the recent results for Gaussian channels to the classical results for generic channels we are able to provide interesting insights into the structure of the capacity region of Gaussian broadcast channels. Probably the most interesting result is in the form of a conjecture towards the end of the chapter. There, it is claimed that all points of the capacity region might be reachable without resorting to time-sharing, i.e., switching between different transmission strategies. This is in contrast to current literature that claims that some points in the capacity region can only be reached by switching between a number of different transmission strategies. This conjecture is shown to be valid for a broadcast channel with two single-antenna receivers in Chapter 1 and for a part of the time-sharing points of a broadcast channel with three single-antenna users in Chapter 2. Avoidance of time-sharing is interesting from a practical point of view as switching between different transmission strategies requires an increased signaling overhead. In this sense, this result, should it be true in all its generality, might be of practical interest. However, it is observed that reaching "time-sharing" points without time-sharing calls for the use of joint encoding, i.e., the information streams of the different users in the network must be encoded jointly rather than successively or independently. Thus, the practical relevance of this result is conditioned on the development of practical joint-encoding approaches.

In Chapter 2 three problems are discussed whose solutions are rate vectors on the boundary of the capacity region. These are the sum-rate maximizing problem, the weighted sum-rate maximizing problem and the rate-balancing problem. For each problem, existing algorithms for memoryless MIMO broadcast channels are reviewed. For the sum-rate and weighted sum-rate maximization problems, it is shown how the block-diagonal structure characteristic of channel matrices in MIMO-OFDM systems can be exploited in order to attain efficient extensions of existing algorithms to time-dispersive channels in some cases, and to develop own algorithmic solutions in others. For the rate-balancing problem, we look at some interesting and significant subtleties around the implementation of optimum subgradient-based approaches and point out some of the shortcomings of this solution such as convergence rate.

The optimum algorithmic solutions discussed in Chapter 2 are all based on an iterative search of the optimum solution. This feature introduces a kind of non-determinism in terms of the computational power required in order to find optimum transmission strategies that is somehow objectionable as far as practical deployment of these algorithms is concerned. That is especially true for scenarios with fast-varying channels, where the quick computa-

tion of the transmit strategy is mandatory in order to leverage the channel state information available at the transmitter. This offers the motivation for considering non-iterative suboptimum approaches in Chapter 3. There, the focus is on what we call decomposition approaches. These are schemes that decompose the broadcast channel into a set of scalar subchannels that are mutually decoupled in the sense that transmission over any particular subchannel does not interfere with transmission over any other subchannel. We first introduce a general framework for decomposition approaches and review some of the existing schemes against this background. Then, a very general algorithm is presented that provides a solution to the general allocation problem of both linear and successive-encoding-based decomposition approaches. In the context of successive-encoding-based schemes, this algorithm includes all other state-of-the-art decomposition algorithms as particular cases. In the context of linear decomposition schemes, the new algorithm represents a holistic approach to the subchannel allocation problem comprising user selection, assignment of spatial dimensions and choice of receive filters. This is in contrast to state-of-the-art algorithms that address these different aspects of the problem separately, following somehow disconnected approaches. The algorithm is specialized to solve each of the problems discussed in Chapter 2 and its performance is evaluated by means of simulations. Performance of the successive-encoding-based decomposition scheme turns out to be almost optimum in all considered scenarios. The performance loss of the linear decomposition scheme, though noticeable, is surprisingly smaller than generally assumed. This result raises some questions regarding the practical relevance of successive encoding approaches, which are notably more difficult to implement. The last section of Chapter 3 deals with a novel non-iterative algorithmic approach to the sum-rate and weighted sum-rate maximization problems that contrary to decomposition approaches results in subchannels with a certain degree of crosstalk. This approach is also based on successive encoding and its performance is observed to be similar to that of the novel successive-encoding-based decomposition algorithm.

Chapter 4 has a different focus than the rest of chapters in this work. It namely deals with the problem of feeding back channel state information (CSI) from the receivers to the transmitter. Different from most of the literature on the topic, which assumes a noiseless or error-free feedback link and considers a constraint on the amount of bits fed back, we adopt a delay-limited paradigm according to which the channel is noisy and the feedback link can be only used a finite number of times in order to transmit CSI. While the rate-limited paradigm can be claimed to be realistic for simple systems where only few bits are needed in order to approach optimality, we find the delay-limited paradigm more convenient for MIMO-OFDM systems. In these systems, due to the relative large amount of information that must be fed back in order to approach optimum performance, nearly error-free transmission is only possible at the cost of significant delay, which might cause the CSI to become obsolete, depending on the rate of variation of the channel. The rate-limited paradigm is also problematic if the feedback channel fades. In such case, regardless delay due to transmission and depending on the amount of diversity in the feedback link, the probability of occurrence of transmission errors might be non-negligible. Sticking to the delay-limited paradigm, first, a SISO-OFDM Rayleigh-fading feedback link is considered and, using mean squared error (MSE) as a figure of merit, some analysis is performed. In particular, upper bounds on performance of transmission over the feedback link are derived based on rate-distortion theory. The tightest bound reveals that, for such a model,

performance is limited either by the degree of diversity available in the feedback link or by the bandwidth expansion of the system, given by the ratio between the dimensionality of the feedback channel and the number of channel coefficients to be fed back. Besides, the optimum linear analog transmission scheme is derived for a case of practical relevance and optimality of linear analog transmission is shown in the low SNR regime. In the high SNR regime, by contrast, linear analog approaches are shown to be unable of leveraging either bandwidth expansion or diversity in terms of distortion decay rate (DDR), which is defined as the asymptotic slope of the function relating MSE and SNR in dB. Based on random codes and a suboptimum maximum-likelihood receiver, we also derive a lower bound on the DDR of digital transmission schemes. This bound reveals that, contrary to linear analog approaches, digital schemes have the potential to exploit both bandwidth expansion and diversity in the high SNR regime. Considering multiple antennas at the receiver of the feedback link we are able to extend some of the results obtained for the SISO-OFDM feedback link and to gain some interesting insights related to the use of multiple antennas. For instance, it is shown that, for a fixed number of subcarriers in the feedback link, a given degree of spectral diversity and under the assumption of uncorrelated antennas, there is an optimum number of antennas that represents the best trade-off between diversity and antenna gain on the one hand, which increase as new antennas are added to the system, and the bandwidth expansion factor, which decreases as new antennas are added to the system due to the increased number of channel coefficients that must be fed back, on the other. This result, which is shown by resorting to theoretical upper bounds on performance, contrasts with the behavior of linear analog transmission whose performance is shown to improve for increasing antenna numbers. That is, linear analog transmission benefits from the increase in antenna gain and diversity to a larger extent than it suffers from the additional burden of channel coefficients to be fed back. The chapter finishes with a discussion on performance measures that are usually utilized in the forward link in order to evaluate the quality of feedback approaches. After questioning the use of ergodic measures in combination with successive-encoding schemes and identifying average throughput as a more suitable measure, simulation results are presented for a simple broadcast channel that suggest that, in spite of its fundamental limitations, a simple linear analog feedback scheme might be good enough to reach a performance close to that achievable when having perfect CSI at the transmitter.

1.2 Notation

Throughout this work, vectors and matrices are denoted by lower case bold and capital bold letters, respectively. Random variables are represented by sans-serif characters. Sets are usually denoted by calligraphic characters. In order to denote a set of indexed elements such as $\{A_i | i = 1, \ldots, I\}$, we frequently use the shortcut $A_{1,\ldots,I}$. For two Hermitian matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} \geq \boldsymbol{B}$ indicates that $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite. A list of frequently used symbols and operators is given in Table 1.1.

$\delta(x)$	Dirac's delta
$\delta[n]$	Kronecker's delta
$\boldsymbol{I}_d$	Identity matrix of dimension $d \times d$
$\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{R})$	Circularly-symmetric complex Gaussian distribution
$\mathbb{R}_+$	Set of non-negative real numbers
$\mathbb{H}^{n\times n}$	Set of Hermitian matrices of dimension $n \times n$
$\lvert \bullet \rvert$	Absolute value of a real or complex number
$\lvert \bullet \rvert$	Determinant of a matrix
$\lvert \bullet \rvert$	Cardinality of a set
$\lVert \bullet \rVert_1$	Manhattan norm
$\lVert \bullet \rVert_2$	Euclidean norm
$[\bullet]_{i,j}$	Entry in row i and column j of a matrix
diag$[\bullet]$	Diagonal matrix with main diagonal defined by the argument
Tr$\{\bullet\}$	Trace operator
E$\{\bullet\}$	Expectation operator
$O(\bullet)$	Big-O of Landau
$o(\bullet)$	Little-o of Landau

Table 1.1: List of frequently used operators and symbols.

1.3 The MIMO OFDM system model

In this work, the channel over which signals propagate from the transmitter to the receiver of a communication link is modeled as a tapped delay line. Assuming L delay taps, t transmit antennas and r receive antennas, this model can be mathematically expressed as

$$\tilde{\boldsymbol{H}}(\tau) = \tilde{\boldsymbol{H}}_1\delta(\tau-\tau_1) + \tilde{\boldsymbol{H}}_2\delta(\tau-\tau_2) + \cdots + \tilde{\boldsymbol{H}}_L\delta(\tau-\tau_L), \tag{1.1}$$

where $\tilde{\boldsymbol{H}}_\ell \in \mathbb{C}^{r\times t}$ is the channel matrix corresponding to the ℓth tap and τ_ℓ is its associated propagation delay. This channel is memoryless if $L = 1$, i.e., the received signal at a particular time instant is just a transformed and possibly delayed version of a signal transmitted at a specific time instant. On the contrary, if $L > 1$, the received signal is, in general, a superposition of signals transmitted at different time instants, i.e., the channel has memory. Such a channel is called time-dispersive. In order to transmit information over this channel, we will consider an orthogonal frequency division multiplexing (OFDM) transmission scheme. At the transmitter, the OFDM symbol has the form

$$\boldsymbol{x}(\xi) = \sum_{n=1}^{N} \boldsymbol{x}_n \mathrm{e}^{\mathrm{j}2\pi f_n \xi} \mathrm{rect}\left(\frac{\xi}{T}\right). \tag{1.2}$$

Here, transmission is performed over N subcarriers. The frequency of subcarrier n is f_n and on this subcarrier a vector $\boldsymbol{x}_n \in \mathbb{C}^t$ is transmitted. The pulse shape, which applies to all transmit antennas, is given by

$$\mathrm{rect}\left(\frac{\xi}{T}\right) = \begin{cases} 1, & \xi \in [0,T] \\ 0, & \xi \notin [0,T] \end{cases},$$

where T is the symbol interval. The subcarrier frequencies are chosen such that $|f_i - f_j| = z/T$, with $z \in \{0, 1, \ldots, N-1\}$. This choice of frequencies makes the subcarriers orthogonal to each other. Using this property, the power of an OFDM symbol can be computed as

$$\frac{1}{T}\int_0^T \|\boldsymbol{x}(\xi)\|_2^2 dt = \sum_{n=1}^{N} \|\boldsymbol{x}_n\|_2^2.$$

In order to avoid intersymbol interference, prior to transmission, a so-called cyclic prefix is appended at the beginning of the symbol. It consists of a replica of a signal block taken from the end of the symbol (cf. Fig. 1.1). After appending the cyclic prefix, Eq. 1.2 becomes

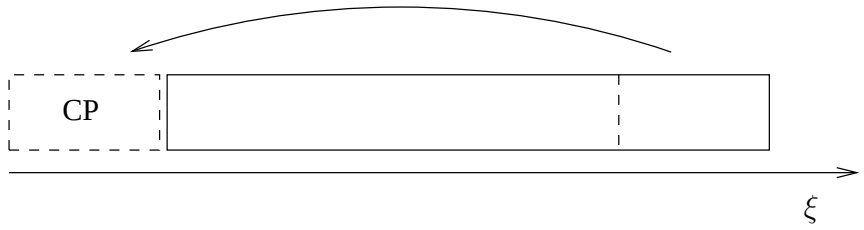

Figure 1.1: Cyclic prefix.

$$\boldsymbol{x}_{\mathrm{CP}}(\xi) = \sum_{n=1}^{N} \boldsymbol{x}_n \mathrm{e}^{\mathrm{j}2\pi f_n \xi} \mathrm{rect}\left(\frac{\xi + T_{\mathrm{CP}}}{T + T_{\mathrm{CP}}}\right), \tag{1.3}$$

where T_{CP} is the duration of the cyclic prefix. The convolution of Eq. 1.1 and Eq. 1.3 yields the expression for the received signal prior to the removal of the cyclic prefix,

$$\tilde{\boldsymbol{y}}_{\mathrm{CP}}(\xi) = \sum_{\ell=1}^{L} \tilde{\boldsymbol{H}}_\ell \sum_{n=1}^{N} \boldsymbol{x}_n \mathrm{e}^{\mathrm{j}2\pi f_n(\xi-\tau_\ell)} \mathrm{rect}\left(\frac{\xi - \tau_\ell + T_{\mathrm{CP}}}{T + T_{\mathrm{CP}}}\right) + \tilde{\boldsymbol{n}}_{\mathrm{CP}}(\xi).$$

Here, $\tilde{\boldsymbol{n}}_{\mathrm{CP}}(\xi) \in \mathbb{C}^r$ is a realization of a multivariate Gaussian stationary stochastic process representing additive noise with zero mean defined in the interval $\xi \in [-T_{\mathrm{CP}} + \tau_1, -T_{\mathrm{CP}} + \tau_L + T]$. As already mentioned, the purpose of the cyclic prefix is to avoid intersymbol interference (ISI) that arises in a multipath channel due to the different delays of the single paths. In order to completely eliminate ISI, its duration must exceed the difference between the shortest and the longest delay in the channel, i.e., $T_{\mathrm{CP}} \geq \tau_L - \tau_1$. If this condition is fulfilled, interference due to the previous symbol remains confined within the interval $[-T_{\mathrm{CP}} + \tau_1, \tau_1]$. That is, it only affects the portion of the symbol corresponding to the cyclic prefix. This also holds for the interference caused by the symbol under consideration on the subsequent symbol. In this case, the interference is confined within the interval $[\tau_1 + T, \tau_L + T]$ which is comprised by the interval $[\tau_1 + T, \tau_1 + T + T_{\mathrm{CP}}]$ corresponding to the cyclic prefix of the subsequent symbol. In order to eliminate intersymbol interference, the cyclic prefix is removed upon reception. The following expression holds for the received symbol after removal of the own cyclic prefix and the cyclic prefix of the subsequent symbol,

$$\tilde{\boldsymbol{y}}(\xi) = \sum_{\ell=1}^{L} \tilde{\boldsymbol{H}}_\ell \sum_{n=1}^{N} \boldsymbol{x}_n \mathrm{e}^{\mathrm{j}2\pi f_n(\xi-\tau_\ell)} \mathrm{rect}\left(\frac{\xi - \tau_1}{T}\right) + \tilde{\boldsymbol{n}}(\xi),$$

where the additive noise is now defined in the interval $[\tau_1, \tau_1+T]$. If this symbol is uniformly sampled at a rate N/T starting at $\xi = \tau_1$ we obtain

$$\begin{aligned}\tilde{\boldsymbol{y}}_m &= \tilde{\boldsymbol{y}}(\tau_1 + (m-1)T/N) = \\ &= \sum_{\ell=1}^{L} \tilde{\boldsymbol{H}}_\ell \sum_{n=1}^{N} \boldsymbol{x}_n \mathrm{e}^{\mathrm{j}2\pi f_n(\tau_1+(m-1)T/N-\tau_\ell)} + \tilde{\boldsymbol{n}}(\tau_1 + (m-1)T/N),\end{aligned}$$

where $m = 1, \ldots, N$. In order to compute the signal received on a particular subcarrier the discrete Fourier transform is applied to these samples as follows,

$$\begin{aligned}\boldsymbol{y}_k &= \frac{1}{N}\sum_{m=1}^{N} \tilde{\boldsymbol{y}}_m \mathrm{e}^{-\mathrm{j}2\pi f_k(m-1)T/N} = \\ &= \sum_{\ell=1}^{L} \tilde{\boldsymbol{H}}_\ell \sum_{n=1}^{N} \boldsymbol{x}_n \mathrm{e}^{\mathrm{j}2\pi f_n(\tau_1-\tau_\ell)} \frac{1}{N}\sum_{m=1}^{N} \mathrm{e}^{\mathrm{j}2\pi(f_n-f_k)(m-1)T/N} + \frac{1}{N}\sum_{m=1}^{N} \tilde{\boldsymbol{n}}_m \mathrm{e}^{\mathrm{j}2\pi f_k(m-1)T/N},\end{aligned}$$

where $\tilde{\boldsymbol{n}}_m = \tilde{\boldsymbol{n}}(\tau_1 + (m-1)T/N)$ and $k = 1, \ldots, N$. Now, noting that

$$\frac{1}{N}\sum_{m=1}^{N} \mathrm{e}^{\mathrm{j}2\pi(f_n-f_k)(m-1)T/N} = \begin{cases} 1, & n = k \\ 0, & n \neq k \end{cases} \tag{1.4}$$

and defining $\boldsymbol{n}_k = \frac{1}{N}\sum_{m=1}^{N} \tilde{\boldsymbol{n}}_m \mathrm{e}^{-\mathrm{j}2\pi f_k(m-1)T/N}$, we can write,

$$\boldsymbol{y}_k = \boldsymbol{H}_k \boldsymbol{x}_k + \boldsymbol{n}_k, \quad k = 1, \ldots, N, \tag{1.5}$$

where $\boldsymbol{H}_k = \sum_{\ell=1}^{L} \tilde{\boldsymbol{H}}_\ell \mathrm{e}^{\mathrm{j}2\pi f_k(\tau_1-\tau_\ell)}$ is the channel matrix corresponding to the kth subcarrier. Since the noise vector in the frequency domain is a linear combination of N Gaussian distributed noise vectors in the time domain, Gaussianity is preserved. That is, $\boldsymbol{n}_k$ is a realization of a multivariate Gaussian distribution, which is also zero-mean. The covariance matrix of this noise is given by

$$\boldsymbol{R} = \mathrm{E}\left\{\boldsymbol{n}_k \boldsymbol{n}_k^{\mathrm{H}}\right\} = \frac{1}{N}\tilde{\boldsymbol{R}},$$

where $\tilde{\boldsymbol{R}} = \mathrm{E}\left\{\tilde{\boldsymbol{n}}_m \tilde{\boldsymbol{n}}_m^{\mathrm{H}}\right\}$ is the covariance matrix of the noise samples in the time domain and it has been assumed that these samples are mutually uncorrelated, i.e., $\mathrm{E}\left\{\tilde{\boldsymbol{n}}_m \tilde{\boldsymbol{n}}_n^{\mathrm{H}}\right\} = \boldsymbol{0}$, $m \neq n$. For most of the discussion in the following chapters, we shall assume $\boldsymbol{R} = \boldsymbol{I}$. Using the assumption of uncorrelated samples in the time domain and Eq. 1.4, it is straightforward to show that noise vectors are also mutually uncorrelated in the frequency domain. Summarizing, as it can be observed in Eq. 1.5, the OFDM transmission scheme transforms the original time-dispersive MIMO channel (cf. Eq. 1.1) into a set of N parallel, memoryless MIMO channels over which information is effectively transmitted.

2 Information Theoretic Fundamentals

2.1 The general broadcast channel

From an information theoretical point of view a system consisting of a transmitter that tries to simultaneously communicate with a number of receivers is a broadcast channel (BC). Broadcast channels were first introduced in [38] and have been extensively discussed in the literature ever since. In this section an overview of key definitions and results on general broadcast channels is given. As it is common use in the literature about the topic, only the two-user case is considered. In general, extensions of definitions and results to broadcast channels with more than two users are trivial. In order to provide the reader with some insight on the theoretical coding schemes that achieve the best performance in some cases and the best known performance in others, the outline of the achievability proofs is given below the corresponding results. These proofs are generally based on the concept of joint typicality and properties of jointly typical sequences. Some background on this topic is given in Appendix A.2.

2.1.1 Definitions

A broadcast channel is a triple $(\mathcal{X}, p(y_1, y_2|x), \mathcal{Y}_1 \times \mathcal{Y}_2)$ consisting of an input alphabet $\mathcal{X}$, two output alphabets $\mathcal{Y}_1$ and $\mathcal{Y}_2$, and a probability transition function $p(y_1, y_2|x)$. Let x^n, y_1^n and y_2^n denote sequences of letters of length n from the alphabets $\mathcal{X}$, $\mathcal{Y}_1$ and $\mathcal{Y}_2$, respectively. The broadcast channel is said to be memoryless if $p(y_1^n, y_2^n|x^n) = \prod_{i=1}^n p(y_{1,i}, y_{2,i}|x_i)$ for any sequence x^n. Given this definition the problem is to know how much information can be sent to both users simultaneously. The transmitter might try to send to both users the same information, like in broadcast television or radio, or might wish to transmit independent information items to each user, as it typically occurs in mobile cellular networks. Simultaneous transmission of independent messages for each user and a common message for both users might also be considered. In the following, we restrict the discussion to the class of memoryless broadcast channels and transmission of independent information.

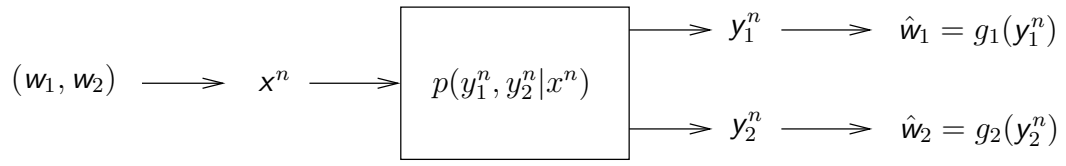

Figure 2.1: Broadcast channel.

For any block length n, a code $\mathcal{C}_n \equiv ((2^{nR_1}, 2^{nR_2}), n)$ for the broadcast channel with independent information for each user consists of a codebook of $2^{n(R_1+R_2)}$ codewords $x^n \in$

$\mathcal{X}^n$, an encoding function ϕ mapping a pair of message indices $(w_1, w_2) \in \{1, \ldots, 2^{nR_1}\} \times \{1, \ldots, 2^{nR_2}\}$ onto the codewords,

$$\phi : \{1, \ldots, 2^{nR_1}\} \times \{1, \ldots, 2^{nR_2}\} \to \mathcal{X}^n,$$

and two decoding functions g_1 and g_2 mapping output sequences onto transmitted messages (see Fig. 2.1),

$$g_1 : \mathcal{Y}_1^n \to \{1, \ldots, 2^{nR_1}\},$$
$$g_2 : \mathcal{Y}_2^n \to \{1, \ldots, 2^{nR_2}\}.$$

Let $(\mathsf{w}_1, \mathsf{w}_2) \in \{1, \ldots, 2^{nR_1}\} \times \{1, \ldots, 2^{nR_2}\}$ be a pair of uniformly distributed random variables. We define the average probability of error as the probability that the decoded messages are not equal to the transmitted messages, i.e.,

$$P_{\mathrm{e}}^{(n)} = P(g_1(y_1^n) \neq \mathsf{w}_1 \text{ or } g_1(y_2^n) \neq w_2).$$

Given this definition of error probability, a rate pair (R_1, R_2) is said to be achievable for the broadcast channel if there exists a sequence of codes $\mathcal{C}_n \equiv ((2^{nR_1}, 2^{nR_2}), n)$ such that $P_{\mathrm{e}}^{(n)} \to 0$ as $n \to \infty$. The capacity region is then defined as the closure of the set of achievable rates. This region is for general broadcast channels still unknown. However, the following theorem can be stated.

Theorem 2.1.1. *The capacity region of the broadcast channel is completely characterized by the conditional marginal distributions $p(y_1|x)$ and $p(y_2|x)$ [40].*

This theorem can be proved by noting that

$$\max\{P_{\mathrm{e},1}^{(n)}, P_{\mathrm{e},2}^{(n)}\} \leq P_{\mathrm{e}}^{(n)} \leq P_{\mathrm{e},1}^{(n)} + P_{\mathrm{e},2}^{(n)},$$

where $P_{\mathrm{e},1}^{(n)} = P(g_1(y_1^n) \neq w_1)$ and $P_{\mathrm{e},2}^{(n)} = P(g_2(y_2^n) \neq w_2)$ are the individual error probabilities of both users. That is, achievability of a certain pair of rates implies that both $P_{\mathrm{e},1}^{(n)} \to 0$ and $P_{\mathrm{e},2}^{(n)} \to 0$ as $n \to \infty$. On the other hand, if both individual error probabilities tend to zero for a given sequence of codes, $P_{\mathrm{e}}^{(n)} \to 0$ and, therefore, the corresponding rates are achievable. Thus, we see that the individual error probabilities completely determine whether a certain pair of rates is achievable or not. Now, the theorem follows from the fact that for a particular sequence of codes the individual error probabilities exclusively depend on the respective conditional marginal distributions.

2.1.2 The degraded broadcast channel

A broadcast channel is said to be physically degraded if x, y_1 and y_2 form a Markov chain $\mathsf{x} \to \mathsf{y}_1 \to \mathsf{y}_2$, i.e., $p(y_1, y_2|x) = p(y_2|y_1)p(y_1|x)$. As an example, we could think of a transmitter sending information over a wired connection to a receiver that, in turn, relays the received signal to a second receiver which is placed farther away in the wired communication path. This second receiver will receive a degraded replica of the signal received by the first. This definition of "degraded" does, in general, not apply to wireless broadcast channels. Indeed, the transmitted signal will, in general, propagate over different

physical paths to reach each of the users, and, hence, no received signal will be a degraded replica of any other. However, still a weaker notion of degraded can be defined that applies to some wireless channels.

A broadcast channel with probability transition function $p(y_1, y_2|x)$ is said to be stochastically degraded if there exists a physically degraded channel with marginal probability transition functions $p(y_1|x)$ and $p(y_2|x)$. Mathematically, this is tantamount to saying that a conditional probability density function $p'(y_2|y_1)$ must exist such that $p(y_2|x) = \sum_{y_1} p'(y_2|y_1)p(y_1|x)$. Note that if a stochastically degraded channel is not physically degraded then $p(y_2|y_1, x) \neq p(y_2|y_1)$.

The capacity region of this kind of broadcast channels is known. It is defined as the convex closure of all rate pairs (R_1, R_2) satisfying

$$\begin{aligned} R_2 &\leq I(\mathsf{u}; \mathsf{y}_2), \\ R_1 &\leq I(\mathsf{x}; \mathsf{y}_1|\mathsf{u}), \end{aligned}$$

for some input distribution $p(u)p(x|u)$. Here, $I(\mathsf{u}; \mathsf{y}_2)$ is the mutual information of the random variables u and y_2, and $I(\mathsf{x}; \mathsf{y}_1|\mathsf{u})$ is the mutual information of x and y_1 conditioned on u [40]. A brief summary of results and curious anecdotes related to this region and the direct and converse proofs can be found in [39]. In the following, the achievability (direct) proof is sketched, which builds upon the interesting concept of superimposed coding.

For given $p(u)$ and $p(x|u)$ a codebook of block length n can be randomly generated as follows. First, generate 2^{nR_2} independent sequences $u^n(w_2)$, $w_2 \in \{1,\ 2,\ \ldots,\ 2^{nR_2}\}$, according to $\prod_{i=1}^{n} p(u_i)$. Then, for each sequence $u^n(w_2)$ generate 2^{nR_1} independent sequences $x^n(w_1, w_2)$, $w_1 \in \{1,\ 2,\ \ldots,\ 2^{nR_1}\}$, according to $\prod_{i=1}^{n} p(x_i|u_i(w_2))$. Before any message is transmitted, the 2^{nR_2} generated $u^n(w_2)$ sequences with their respective indexes are revealed to user 2. These indexed sequences are also revealed to user 1 together with the $2^{n(R_2+R_1)}$ indexed sequences $x^n(w_1, w_2)$.

In order to transmit a message w_1 to user 1 and a message w_2 to user 2, the encoder selects the corresponding sequence $x^n(w_1, w_2)$, which is transmitted over the channel. Let y_2^n be the sequence received by user 2. In order to find out the message w_2 that was transmitted, this user looks for a codeword u^n in the codebook so that this sequence and y_2^n are jointly typical. For large n, such a sequence will exist almost surely and it will be unique, with high probability, provided that $R_2 \leq I(\mathsf{u}; \mathsf{y}_2) - \epsilon$, $\epsilon > 0$. In the limit $n \to \infty$, this condition allows user 2 to reliably retrieve the transmitted message w_2 out of the received sequence. Let y_1^n be the sequence received by user 1. Due to the degraded quality of the channel $I(\mathsf{u}; \mathsf{y}_1) \geq I(\mathsf{u}; \mathsf{y}_2)$[1] and, therefore, user 1 is in a position to also identify the transmitted w_2 reliably. Once this is done, this user looks among the 2^{nR_1} codewords x^n associated with w_2 for a sequence that is jointly typical with y_1^n. In this case, reliable detection of the transmitted message w_1 is possible for $n \to \infty$ provided that $R_1 \leq I(\mathsf{x}; \mathsf{y}_1|\mathsf{u})$. Thus, capacity is achieved by coding information in two layers. A coarse layer represented by the sequences u^n that can be detected by both users and a second layer of finer information upon the first layer that is represented by the sequences x^n and can be only perceived by the best user. Note that sequences u^n are not explicitly

[1] Data processing inequality [40].

transmitted. However, any transmitted signal x^n reveals the identity of the sequence u^n from which it was generated.

2.1.3 The non-degraded broadcast channel

The capacity region of the general non-degraded broadcast channel is still unknown. Exceptions are broadcast channels with deterministic components and the Gaussian broadcast channel. For the latter, the capacity region has been recently found and will be discussed in Section 2.2. The capacity region of deterministic broadcast channels was found in the late seventies (see [39] and references therein). A generalization of this result to arbitrary broadcast channels gave rise to the largest achievability region for general broadcast channels known so far, which turns out to be the capacity region if the broadcast channel has a deterministic component [80].

2.1.3.1 Marton's achievability region

According to Marton's result in [80], for the case in which only independent information is transmitted to the users, the rates (R_1, R_2) are achievable for the broadcast channel $(\mathcal{X}, p(y_1, y_2|x), \mathcal{Y}_1 \times \mathcal{Y}_2)$ if

$$\begin{aligned} R_1 &\leq I(u_1; y_1), && (2.1)\\ R_2 &\leq I(u_2; y_2), && (2.2)\\ R_1 + R_2 &\leq I(u_1; y_1) + I(u_2; y_2) - I(u_1; u_2), && (2.3) \end{aligned}$$

for some $p(x, u_1, u_2)$ on $\mathcal{X} \times \mathcal{U}_1 \times \mathcal{U}_2$.

A simple proof of the achievability of this region given in [45] goes as follows (see Fig. 2.2). In the first step, $2^{nI(u_1;y_1)}$ sequences u_1^n and $2^{nI(u_2;y_2)}$ sequences u_2^n are generated according to distributions $p(u_1)$ and $p(u_2)$, respectively. Sequences u_1^n are uniformly distributed over 2^{nR_1} bins and sequences u_2^n are uniformly distributed over 2^{nR_2} bins. Each pair of message indices (w_1, w_2) identifies a bin in the grid of $2^{nR_1} \times 2^{nR_2}$ bins. The correspondence between message index w_1 and the 2^{nR_1} bins containing sequences u_1^n is known to receiver 1 and the correspondence between message index w_2 and the 2^{nR_2} bins containing sequences u_2^n is known to receiver 2.

In the second step, given a pair of message indices (w_1, w_2), a pair of jointly typical sequences (u_1^n, u_2^n) is taken from the corresponding bin. The probability that two sequences u_1^n and u_2^n taken at random are jointly typical is given by $2^{-nI(u_1;u_2)}$. Hence, at least $2^{nI(u_1;u_2)}$ pairs of sequences (u_1^n, u_2^n) per bin are needed so that the existence of jointly typical pairs is guaranteed with probability of almost 1 for large n. From this and from the fact that there are a number of $2^{n(I(u_1;y_1)+I(u_2;y_2)-R_1-R_2)}$ pairs of sequences per bin, the condition

$$R_1 + R_2 \leq I(u_1; y_1) + I(u_2; y_2) - I(u_2; u_1)$$

results that must be satisfied in order to assure the integrity of the encoding procedure.

In the third encoding step the sequence x^n of transmit signals is drawn according to the distribution given by $p(x^n|u_1^n, u_2^n)$. Upon receiving the sequence y_1^n, receiver 1 looks for

a sequence u_1^n in the codebook which is jointly typical with the received sequence. This sequence will be unique with probability of almost 1 provided that

$$R_1 \leq I(u_1; y_1).$$

In a similar way receiver 2 requires

$$R_2 \leq I(u_2; y_2)$$

in order to be able to detect the transmitted sequence u_2^n reliably. Finally, the receivers map the bins containing the detected sequences u_1^n and u_2^n to the corresponding message indices $\hat{w}_1$ and $\hat{w}_2$.

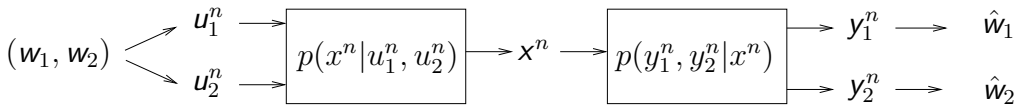

Figure 2.2: Broadcast coding.

In the description above, encoding of both messages happens simultaneously. Alternatively, these rates can also be achieved by encoding the messages to be transmitted not simultaneously but successively. While the first message is encoded without considering information intended for the second user, the encoding of the second message regards the already encoded message for the first user as interference, which, of course, is known at the transmitter. In fact, this is basically the approach taken in [80] in order to prove achievability. This was already observed by Gelfand and Pinsker in [54], where they derived the capacity of a general discrete memoryless channel with non-causally known interference at the transmitter.

2.1.3.2 Coding with known interference and Marton's region

Let $(\mathcal{S}, p(y|s,u), \mathcal{Y})$ be a discrete memoryless channel for which the probability of a certain output sequence y^n depends not only on the particular input sequence s^n but also on an interfering sequence u^n that is known to the transmitter but unknown to the receiver. For some fixed $p(u,v,s)$ defined on $\mathcal{U} \times \mathcal{V} \times \mathcal{S}$, the maximum rate achievable over this channel is given by

$$R = I(v; y) - I(v; u),$$

where v is an auxiliary random variable selected from a finite alphabet $\mathcal{V}$ [54]. Capacity can be achieved by coding as follows (see Fig. 2.3).

First, $2^{nI(v;y)}$ sequences v^n are drawn according to $p(v)$ and are uniformly distributed over 2^{nR} bins. Every bin is assigned a message index $w \in \{1, \ldots, 2^{nR}\}$. The mapping of message indices to bins and the correspondence between sequences v^n and bins are known to the receiver.

Secondly, given an interfering sequence u^n and a message index to be transmitted, in the corresponding bin a sequence v^n is looked for that is jointly typical with the given sequence

u^n. Using similar arguments as in the previous section, it can be shown that the existence of a jointly typical v^n in each bin given a typical interference sequence u^n requires

$$R \leq I(v; y) - I(v; u).$$

In the final step of the encoding procedure the input sequence s^n is generated according to $p(s^n|v^n, u^n)$. Upon receiving a sequence y^n, the receiver looks in the codebook for a jointly typical sequence v^n which, for large n, will be unique with probability of almost 1 provided that the number of sequences v^n in the codebook is not larger than $2^{nI(v;y)}$. Finally, the message index $\hat{w}$ is identified that corresponds to the bin containing the detected sequence v^n.

Note that the interference has an impact on the output of the channel and is taken into account during generation of the codebook and the transmit signal. This is indicated by the three arrows departing from u^n in Fig. 2.3.

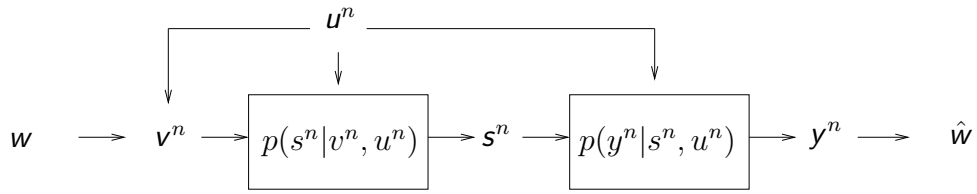

Figure 2.3: Coding with known interference.

As already mentioned, the described coding strategy with known interference can be applied to broadcast channels if a successive encoding approach is chosen (see Fig. 2.4).

Given a broadcast channel $(\mathcal{X}, p(y_1, y_2|x), \mathcal{Y}_1 \times \mathcal{Y}_2)$ and a distribution $p(x, s, u_1, u_2)$ defined on $\mathcal{X} \times \mathcal{S} \times \mathcal{U}_1 \times \mathcal{U}_2$ we can proceed by coding information for user 1 first without considering user 2 at all. In this case, a transmission free of errors for user 1 can only be achieved if the corresponding codebook contains less than $2^{nI(u_1;y_1)}$ sequences u_1^n, i.e., Eq. 2.1 must be satisfied.

For user 2, coding can be done according to the strategy described above considering any coded sequence u_1^n for user 1 as interference in the transmission channel to user 2. Correspondingly, an error-free transmission for user 2 can only be achieved if

$$R_2 \leq I(u_2; y_2) - I(u_2; u_1).$$

This inequality together with the rate limit for user 1 results in the inequality for the sum of rates given by Eq. 2.3. If the coding steps are inverted, i.e., we first code information for user 2 and then information for user 1, considering the signal transmitted to user 2 as known interference, Eq. 2.2 holds for R_2 and

$$R_1 \leq I(u_1; y_1) - I(u_1; u_2),$$

which together with Eq. 2.2 results in Eq. 2.3. Thus, we observe that the limits of Marton's region can also be achieved by applying a successive encoding approach.

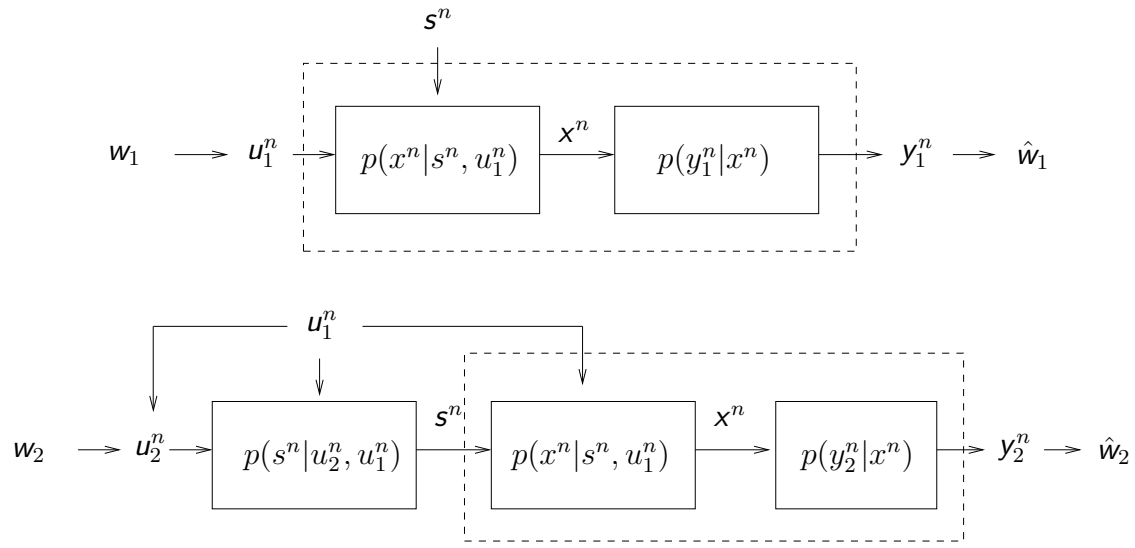

Figure 2.4: Successive coding in broadcast channels.

Fig. 2.4 shows the relationship between the different signals involved in the encoding process. In contrast to the simultaneous encoding approach illustrated in Fig. 2.2, the successive encoding scheme requires the introduction of a variable s that represents the signal which is sent by user 2 and that is used together with signal u_1, issued by user 1, for the generation of signal x, which is actually transmitted over the channel.

2.1.3.3 Marton's region and degraded channels

If applied to a degraded broadcast channel, Marton's region must lie within the boundaries of the capacity region discussed in Section 2.1.2. However, the question arises whether the capacity region is strictly larger than Marton's region in this case. The answer is yes unless some properties hold for the signals involved in the encoding process.

Assume that the degraded broadcast channel $(\mathcal{X}, p(y_1, y_2|x), \mathcal{Y}_1 \times \mathcal{Y}_2)$, $x \to y_1 \to y_2$, is given and the codebook of signals intended for user 2 is generated from an alphabet $\mathcal{U}_2$ according to a distribution $p(u_2)$. The transmitted signals are generated according to a distribution $p(x|u_2)$. We recall that the capacity region for this setting is given by the following inequalities,

$$\begin{aligned} R_2 &\leq I(u_2; y_2), \\ R_1 &\leq I(x; y_1|u_2). \end{aligned}$$

Fix $p(x, u_2)$ and consider an alphabet $\mathcal{U}_1$ of signals intended for the first user and the joint probability density function $p(x, u_2, u_1)$. According to Marton's result, the following rates are achievable

$$\begin{aligned} R_2 &\leq I(u_2; y_2), \\ R_1 &\leq I(u_1; y_1) - I(u_1; u_2). \end{aligned}$$

While the achievable rate for user 2 is bounded by the same limit in both regions, the limits for user 1 are different. Indeed, the following relations hold,

$$\begin{aligned} I(\mathsf{x};\mathsf{y}_1|\mathsf{u}_2) &\geq I(\mathsf{u}_1;\mathsf{y}_1|\mathsf{u}_2) \\ &= I(\mathsf{u}_1;\mathsf{y}_1,\mathsf{u}_2) - I(\mathsf{u}_1;\mathsf{u}_2) \\ &= I(\mathsf{u}_1;\mathsf{y}_1) + I(\mathsf{u}_1;\mathsf{u}_2|\mathsf{y}_1) - I(\mathsf{u}_1;\mathsf{u}_2) \\ &\geq I(\mathsf{u}_1;\mathsf{y}_1) - I(\mathsf{u}_1;\mathsf{u}_2). \end{aligned} \tag{2.4}$$

The first inequality is obtained by observing that the variables u_1, x and y_1 form a Markov chain $\mathsf{u}_1 \rightarrow \mathsf{x} \rightarrow \mathsf{y}_1$ and applying the data-processing inequality [40]. The last inequality results from the fact that mutual information is always non-negative. The equalities are simple applications of the chain rule for mutual information. As we expected the capacity region is larger or equal to Marton's region. Still, equality can be achieved in the above inequalities if it is possible to choose the variable u_1 such that the transmitted signal x is a deterministic function of u_2 and u_1, i.e., $x = f(u_1, u_2)$, and $I(\mathsf{u}_1;\mathsf{y}_1|\mathsf{u}_2) = I(\mathsf{u}_1;\mathsf{y}_1) - I(\mathsf{u}_1;\mathsf{u}_2)$. The first condition leads to equality in Eq. 2.4. The second condition implies that the rate achieved over a channel where the interference is only known at the transmitter must be the same as that achieved when the interference is known at the receiver. We shall see that both conditions can be fulfilled if the broadcast channel is Gaussian.

2.1.3.4 Sato bound

Marton's achievability region represents an inner bound to the capacity region of arbitrary broadcast channels. A well known outer bound to the capacity region of arbitrary broadcast channels was presented by Sato in [99]. This result states that for a broadcast channel $(\mathcal{X}, p(y_1, y_2|x), \mathcal{Y}_1 \times \mathcal{Y}_2)$ the capacity region for a given choice of $p(x)$ is confined within the following region,

$$\begin{aligned} R_1 &\leq I(\mathsf{x};\mathsf{y}_1), \\ R_2 &\leq I(\mathsf{x};\mathsf{y}_2), \\ R_1 + R_2 &\leq \min_{p(y_1,y_2|x)} \{I(\mathsf{x};\mathsf{y}_1,\mathsf{y}_2)\}, \text{ subject to given } p(y_1|x) \text{ and } p(y_2|x). \end{aligned} \tag{2.5}$$

The bounds for the individual rates are trivial and were already noted by Cover in [38]. The bound on the sum rate relies on two simple facts. The first is that $I(\mathsf{x};\mathsf{y}_1,\mathsf{y}_2)$ represents the maximum throughput achievable over the channel if the two users can cooperate. If the users can not cooperate the throughput will be necessarily less or equal. Indeed, users with cooperation capability might always choose not to cooperate should this strategy increase the sum rate. The second fact is that the capacity region of the broadcast channel does not depend on the joint transition probability function $p(y_1, y_2|x)$ but only on the marginal distributions $p(y_1|x)$ and $p(y_2|x)$ (see Theorem 2.1.1). As $I(\mathsf{x};\mathsf{y}_1,\mathsf{y}_2)$ does depend on $p(y_1, y_2|x)$ the tightness of the cooperative bound can be maximized by minimizing over the joint transition probability function while keeping the marginals fixed.

2.2 The Gaussian broadcast channel

Having reviewed basic information theoretic results on general broadcast channels, we now turn our attention to Gaussian broadcast channels. A broadcast channel $(\mathcal{X}, p(y_1, y_2, \ldots, y_K|x), \mathcal{Y}_1 \times \mathcal{Y}_2 \times \cdots \times \mathcal{Y}_K)$ is said to be Gaussian if the probability transition function has the form of a multivariate Gaussian probability density function, being the input and the output alphabets real Euclidean spaces. Alternatively, if the physical medium permits transmission of in-phase and quadrature components, complex Euclidean spaces can be considered as input and output alphabets and the probability transition function adopts the form of a multivariate circularly symmetric complex Gaussian probability density function [119]. In the following, this last class of Gaussian broadcast channels will be considered.

Assume a t-dimensional complex-valued input alphabet, i.e., $\mathcal{X} = \mathbb{C}^t$, and r_k-dimensional complex-valued output alphabets, $\mathcal{Y}_k = \mathbb{C}^{r_k}$. Gaussian broadcast channels admit the following algebraic representation,

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n} \tag{2.6}$$

with

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_K \end{bmatrix}, \quad \boldsymbol{n} = \begin{bmatrix} \boldsymbol{n}_1 \\ \vdots \\ \boldsymbol{n}_K \end{bmatrix}, \quad \boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_1 \\ \vdots \\ \boldsymbol{H}_K \end{bmatrix}.$$

In this model, $\boldsymbol{H}_k \in \mathbb{C}^{r_k \times t}$, which is called the channel matrix of user k, describes the transformation that the transmitted signal $\boldsymbol{x} \in \mathbb{C}^t$ experiences while propagating from the transmitter to receiver k. The resulting signal is corrupted by an additive circularly symmetric complex Gaussian noise vector $\boldsymbol{n}_k \in \mathbb{C}^{r_k}$ giving rise to the received signal $\boldsymbol{y}_k \in \mathbb{C}^{r_k}$. Usually, the noise is assumed to be zero-mean and a constraint is considered that limits the average power of the transmitted signal, i.e., $\mathrm{E}\{\|\boldsymbol{x}\|_2^2\} \leq P$.

2.2.1 The single-input Gaussian broadcast channel

In this section we examine the single-input, $t = 1$, Gaussian broadcast channel. A physical scenario corresponding to this model is the downlink of a cellular communication system with a single antenna at the transmitter (see Fig. 2.5). For the two user case, i.e., K=2, the channel is given by a pair of equations,

$$\boldsymbol{y}_1 = \boldsymbol{h}_1 x + \boldsymbol{n}_1,$$

$$\boldsymbol{y}_2 = \boldsymbol{h}_2 x + \boldsymbol{n}_2.$$

In the following, it is shown that this is a degraded broadcast channel. To this end, first, consider the broadcast channel obtained by applying matched filters at the receivers,

$$\tilde{y}_1 = x + \tilde{n}_1, \tag{2.7}$$

$$\tilde{y}_2 = x + \tilde{n}_2, \tag{2.8}$$

where $\tilde{y}_k = \alpha_k^{-1} \boldsymbol{h}_k^{\mathrm{H}} \boldsymbol{R}_k^{-1} \boldsymbol{y}_k$, being $\boldsymbol{R}_k$ the covariance matrix of the noise process affecting user k and $\alpha_k = \boldsymbol{h}_k^{\mathrm{H}} \boldsymbol{R}_k^{-1} \boldsymbol{h}_k$.

It can be shown that this processing at the receivers preserves capacity and, therefore, Eqs. 2.7, 2.8 constitute an alternative model for the original channel. In order to see that this is true, consider a family of codes $\mathcal{C}_n \equiv ((2^{nR_1}, 2^{nR_2}), n)$ such that transmission errors tend to zero as $n \to \infty$ for an achievable pair of rates (R_1, R_2). Without loss of optimality, the decoders can be assumed to be maximum-likelihood estimators,

$$g_k(\boldsymbol{y}_k^n) = \underset{1 \leq w_k \leq 2^{nR_k}}{\arg\max} \{p(\boldsymbol{y}_k^n | w_k)\} = \hat{w}_k.$$

Note that error-free transmission implies that the receivers are able to recognize unambiguously the message that originated the received sequence and, therefore, as $n \to \infty$, $p(\boldsymbol{y}_k^n | w_k) \to 0$ for all except for the transmitted message. Now, we shall show that $p(\boldsymbol{y}_k^n | w_k) = f(\boldsymbol{y}_k^n) p(\tilde{y}_k^n | w_k)$, i.e., the maximizer of $p(\boldsymbol{y}_k^n | w_k)$ is also the maximizer of $p(\tilde{y}_k^n | w_k)$ and, therefore, $\tilde{y}_k^n$ is a sufficient statistic for detection of the transmitted message [67]. The likelihood function $p(\boldsymbol{y}_k^n | w_k)$ can be written in terms of the transmitted sequence x^n as follows,

$$\begin{aligned} p(\boldsymbol{y}_k^n | w_k) &= \int_{\mathbb{C}^n} p(\boldsymbol{y}_k^n | x^n, w_k) p(x^n | w_k) dx^n, \\ &= \int_{\mathbb{C}^n} p(\boldsymbol{y}_k^n | x^n) p(x^n | w_k) dx^n, \end{aligned} \tag{2.9}$$

where the second equality is due to the fact that $\mathsf{w}_k \to \mathsf{x} \to \mathsf{y}_k$ form a Markov chain. In turn, using the fact that the channel is memoryless and some algebra, we can write

$$\begin{aligned} p(\boldsymbol{y}_k^n | x^n) &= \prod_{i=1}^{n} p(\boldsymbol{y}_{k,i} | x_i), \\ &= \prod_{i=1}^{n} \frac{1}{\pi^{r_k} |\boldsymbol{R}_k|} \exp -(\boldsymbol{y}_{k,i} - \boldsymbol{h}_k x_i)^{\mathrm{H}} \boldsymbol{R}_k^{-1} (\boldsymbol{y}_{k,i} - \boldsymbol{h}_k x_i), \\ &= \prod_{i=1}^{n} \frac{1}{\pi^{r_k} |\boldsymbol{R}_k|} \exp - \left(\boldsymbol{y}_{k,i}^{\mathrm{H}} \boldsymbol{R}_k^{-1} \boldsymbol{y}_{k,i} - \alpha_k |\tilde{y}_{k,i}|^2 \right) \exp -\alpha_k |\tilde{y}_{k,i} - x_i|^2, \\ &= f(\boldsymbol{y}_k^n) \prod_{i=1}^{n} \frac{\alpha_k}{\pi} \exp -\alpha_k |\tilde{y}_{k,i} - x_i|^2, \\ &= f(\boldsymbol{y}_k^n) p(\tilde{y}_k^n | x^n). \end{aligned}$$

Now, plugging this result into Eq. 2.9 we finally obtain

$$\begin{aligned} p(\boldsymbol{y}_k^n | w_k) &= f(\boldsymbol{y}_k^n) \int_{\mathbb{C}^n} p(\tilde{y}_k^n | x^n) p(x^n | w_k) dx^n, \\ &= f(\boldsymbol{y}_k^n) \int_{\mathbb{C}^n} p(\tilde{y}_k^n | x^n, w_k) p(x^n | w_k) dx^n, \\ &= f(\boldsymbol{y}_k^n) p(\tilde{y}_k^n | w_k). \end{aligned}$$

Let σ_1^2 and σ_2^2 be the variances of the zero-mean Gaussian variables $\tilde{n}_1$ and $\tilde{n}_2$, respectively, and assume that $\sigma_2^2 \geq \sigma_1^2$, without loss of generality. Then, the following physically

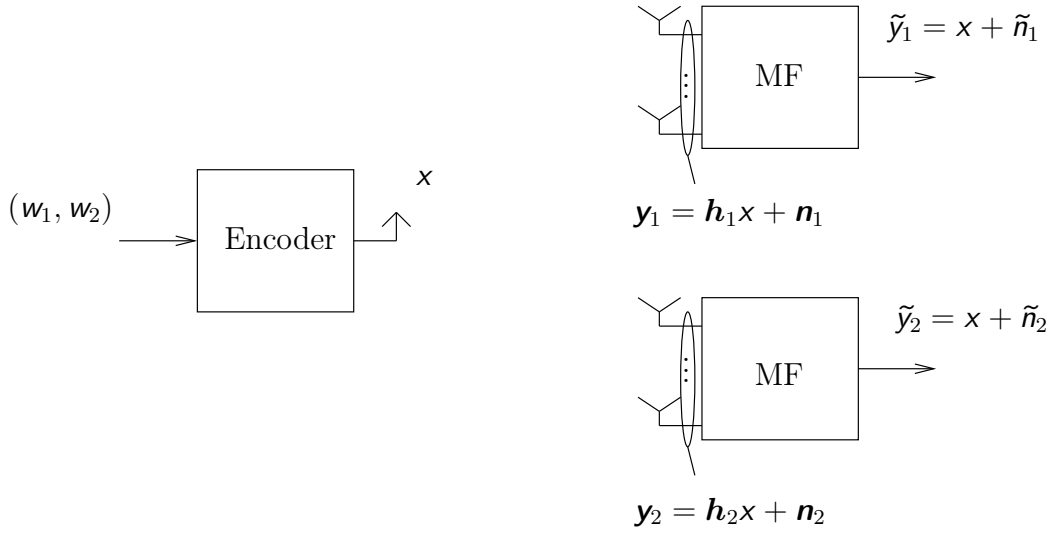

Figure 2.5: Gaussian broadcast channel with single transmit antenna.

degraded broadcast channel can be defined

$$\bar{y}_1 = x + \tilde{n}_1,$$

$$\bar{y}_2 = x + \tilde{n}_1 + \bar{n}_2,$$

with $\mathrm{E}\{|\bar{n}_2|^2\} = \sigma_2^2 - \sigma_1^2$. While in general $p(\bar{y}_1, \bar{y}_2|x) \neq p(\tilde{y}_1, \tilde{y}_2|x)$, it is easy to see that $p(\bar{y}_1|x) = p(\tilde{y}_1|x)$ and $p(\bar{y}_2|x) = p(\tilde{y}_2|x)$. Thus, we note that for every Gaussian broadcast channel with a single transmit antenna an equivalent physically degraded broadcast channel can be found, which proves that these channels are themselves degraded.

In [5], Bergmans showed that, under an average power constraint, all achievable rate pairs can be reached by using a Gaussian distribution for the generation of the codebook. Using the notation employed in Section 2.1.2, the signals intended for user 2 are generated according to $\boldsymbol{u} \sim \mathcal{CN}(0, (1-\alpha)P)$ and

$$p(x|u) = \frac{1}{\pi\alpha P} \exp -\frac{|x-u|^2}{\alpha P}$$

for $0 \leq \alpha \leq 1$.[2] Equivalently, we can denote the signal intended for user 2 by $u_2 = u$ and define a signal u_1 that is intended for the first user, statistically independent with respect to $\boldsymbol{u}_2$ and generated according to $\boldsymbol{u}_1 \sim \mathcal{CN}(0, \alpha P)$. The transmitted signal is given by

[2] In [5] a broadcast channel with real-valued inputs and outputs is assumed. The results obtained under this assumption immediately apply to the case of complex-valued inputs and outputs by noting that, due to the circularly symmetric noise, the channel given by Eqs. 2.7 and 2.8 can be viewed as a superposition of two parallel real-valued broadcast channels.

$x = u_1 + u_2$. Using this notation, we can see that the pairs of achievable rates are given by

$$R_1 = \log\left(1 + \frac{\alpha P}{\sigma_1^2}\right), \tag{2.10}$$

$$R_2 = \log\left(1 + \frac{(1-\alpha)P}{\sigma_2^2 + \alpha P}\right). \tag{2.11}$$

The boundary of the capacity region, defined by these equations and parameterized by α, can be visualized in Fig. 2.6 for a broadcast channel with $P = 10$, $\sigma_1^2 = 1/2$ and $\sigma_2^2 = 1$. A salient property of degraded Gaussian broadcast channels is that the sum capacity is always achieved by assigning all power to the best user. This is easily shown by adding Eqs. 2.10 and 2.11 and rewriting the resulting expression as

$$R_1 + R_2 = \log\left(\sigma_1^2 + \alpha P\right) - \log\left(\sigma_2^2 + \alpha P\right) + \log\left(\frac{\sigma_2^2 + P}{\sigma_1^2}\right). \tag{2.12}$$

This expression reaches its maximum at $\alpha = 1$.

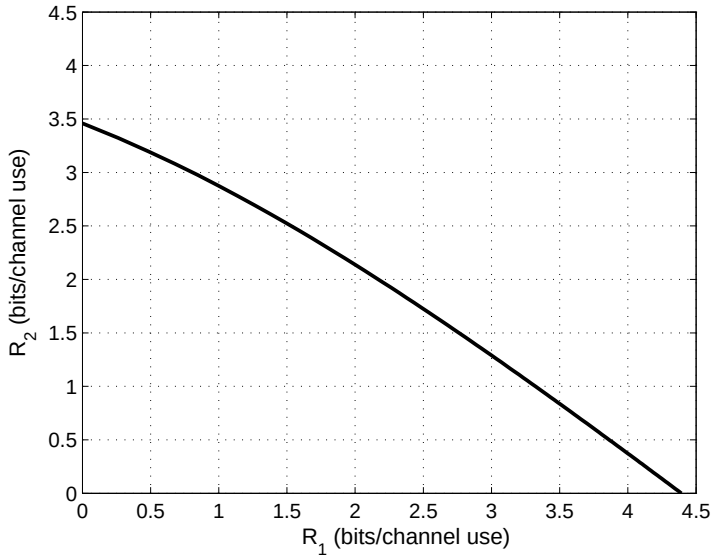

Figure 2.6: Capacity region for a degraded Gaussian broadcast channel with $P = 10$ dB, $\sigma_1^2 = 1/2$ and $\sigma_2^2 = 1$.

2.2.2 The multiple-input Gaussian broadcast channel

In contrast to single-input Gaussian broadcast channels, multiple-input Gaussian broadcast channels are, in general, non-degraded. In order to illustrate this fact consider the two-user channel given by

$$\boldsymbol{y}_1 = \boldsymbol{H}_1\boldsymbol{x} + \boldsymbol{n}_1,$$

$$\boldsymbol{y}_2 = \boldsymbol{H}_2\boldsymbol{x} + \boldsymbol{n}_2,$$

with $\boldsymbol{x} \in \mathbb{C}^t$, $t > 1$, and assume that there exists $\boldsymbol{x}_0 \neq \boldsymbol{0}$ such that $\boldsymbol{x}_0 \in \text{Null}\{\boldsymbol{H}_1\}$ and $\boldsymbol{x}_0 \notin \text{Null}\{\boldsymbol{H}_2\}$. If this channel were degraded, a conditional probability density function $p'(\boldsymbol{y}_2|\boldsymbol{y}_1)$ would exist such that

$$p(\boldsymbol{y}_2|\boldsymbol{x}) = \int_{\mathbb{C}^{r_1}} p'(\boldsymbol{y}_2|\boldsymbol{y}_1)p(\boldsymbol{y}_1|\boldsymbol{x})d\boldsymbol{y}_1 \tag{2.13}$$

for all values of $\boldsymbol{x}$. For any input $\alpha\boldsymbol{x}_0$, we observe that $p(\boldsymbol{y}_1|\alpha\boldsymbol{x}_0) = p(\boldsymbol{n}_1)$ and, therefore, for any fixed $p'(\boldsymbol{y}_2|\boldsymbol{y}_1)$ the right-hand side of Eq. 2.13 is not a function of the scalar α. However, the left-hand side clearly depends on α. Thus, no $p'(\boldsymbol{y}_2|\boldsymbol{y}_1)$ can be found that satisfies Eq. 2.13 for all possible inputs and, therefore, the channel is not degraded.

Even if $\text{Null}\{\boldsymbol{H}_1\} = \text{Null}\{\boldsymbol{H}_2\}$ the multiple-input channel is generally non-degraded. This is due to the fact that, in general, a ranking of vector channels can not be established. As an example, consider two users with $\boldsymbol{H}_1 = \boldsymbol{H}_2 = \boldsymbol{I}_d$, $d > 1$, and noise vectors $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ such that $\text{E}\{\boldsymbol{n}_1\boldsymbol{n}_1^{\text{H}}\} = \text{diag}[\sigma_1^2, \cdots, \sigma_d^2]$ and $\text{E}\{\boldsymbol{n}_2\boldsymbol{n}_2^{\text{H}}\} = \text{diag}[\sigma_d^2, \cdots, \sigma_1^2]$, with $\sigma_1^2 > \cdots > \sigma_d^2$. By no means can a physically degraded model be found according to which the signal received by one user is a degraded replica of the signal that the other user receives. This happens because it is impossible to establish an order between the noise vectors experienced by the users.

In the next sections, first, a particular instance of the Marton region is discussed for this kind of channels. This region is based on successive encoding and a particularization of the coding scheme presented in [54] to Gaussian channels called dirty paper coding (DPC) [37]. Then, we will describe some of the most important properties of this region such as the duality with the capacity region of the multiple access channel (MAC). Finally, we will conclude with a brief review of the work that led to the finding that this region is actually the capacity region of the Gaussian channel with multiple inputs.

2.2.2.1 Writing on Dirty Paper

The title of this section is the same as the title of the famous paper by Costa [37]. It emphasizes the analogy between writing on a paper with spots or dirt and coding on a channel with known interference at the transmitter. The writer, who intends to cipher a message for the reader, knows where the spots are placed on the paper and uses this knowledge to counteract any adverse effect of these on the communication process. Similarly, the encoder knows the interference and applies such knowledge so as to counteract its negative effects. Both reader and receiver can perfectly ignore the interference while being able to decode the message. That means that the receiver does not need to know the interference in order to recover the transmitted message.

Essentially, in [37], Costa applies the result presented in [54] and discussed in Section 2.2.2.2 to the scalar Gaussian channel and derives the optimum distribution for the input parameters.[3]

For the channel

$$y = s + u + z$$

[3]Note that, though derived for discrete alphabets, the result presented in [54] and discussed in Section 2.2.2.2 can be readily extended to continuous alphabets by using the methods employed in [52] in order to prove the channel coding theorem for continuous alphabets.

with noise $\mathsf{z} \sim \mathcal{CN}(0, \sigma_z^2)$, interference $\mathsf{u} \sim \mathcal{CN}(0, \sigma_u^2)$ and transmit power constraint $\mathrm{E}\{|\mathsf{s}|^2\} \leq P$, Costa showed that optimality is achieved by choosing a Gaussian probability density function $p(v, u, s)$ with zero mean and covariance matrix[4]

$$\boldsymbol{R} = \mathrm{E}\left\{ \begin{bmatrix} \mathsf{v} \\ \mathsf{u} \\ \mathsf{s} \end{bmatrix} \begin{bmatrix} \mathsf{v}^* & \mathsf{u}^* & \mathsf{s}^* \end{bmatrix} \right\} = \begin{bmatrix} P + \alpha^2\sigma_u^2 & \alpha\sigma_u^2 & P \\ \alpha\sigma_u^2 & \sigma_u^2 & 0 \\ P & 0 & P \end{bmatrix},$$

with

$$\alpha = \frac{P}{P + \sigma_z^2}.$$

Furthermore, the capacity thus achieved was shown to be the same as that of the Gaussian channel without the known interference, i.e.,

$$I(s; y|u) = I(v; y) - I(v; u). \tag{2.14}$$

Thus, in such a channel interference does not diminish capacity as long as it is known at the transmitter. Note that this optimum coding scheme results in transmit signals and interference being mutually uncorrelated.

In [140], this result was extended to vector channels,

$$\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{u} + \boldsymbol{z}, \tag{2.15}$$

with noise $\boldsymbol{z} \sim \mathcal{CN}(0, \boldsymbol{R}_z)$, interference $\boldsymbol{u} \sim \mathcal{CN}(0, \boldsymbol{R}_u)$ and transmit power constraint $\mathrm{E}\{\|\boldsymbol{s}\|_2^2\} \leq P$. Optimality is also achieved by choosing the transmit signal $\boldsymbol{s}$ to be statistically independent of the interference $\boldsymbol{u}$ and $\boldsymbol{v} = \boldsymbol{s} + \boldsymbol{\Gamma}\boldsymbol{u}$ with $\boldsymbol{\Gamma} = \boldsymbol{R}_s \left(\boldsymbol{R}_s + \boldsymbol{R}_z\right)^{-1}$, where $\boldsymbol{R}_s = \mathrm{E}\{\boldsymbol{s}\boldsymbol{s}^{\mathrm{H}}\}$. The resulting rate is the same as that achievable over the channel $\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{z}$ with transmit covariance matrix $\boldsymbol{R}_s$.

In recent years there have been further extensions of this result in several directions. In [148] the authors show that the known interference can also be completely neutralized even if the interference and noise processes are not stationary or ergodic. In [35] the authors show that this result also holds for an ergodic known interference with arbitrary distribution provided that the noise is Gaussian distributed. Finally, in [46] it is shown that the result also holds for an arbitrarily varying known interference provided that common randomness is shared by transmitter and receiver.

2.2.2.2 Dirty paper coding region

Given a Gaussian broadcast channel as represented by Eq. 2.6 and considering the Marton achievability region discussed in cf. Section 2.1.3.1, the question arises about how to choose the joint probability density function $p(\boldsymbol{x}, \boldsymbol{u}_1, \cdots, \boldsymbol{u}_K)$ of the auxiliary variables $\boldsymbol{u}_k$, $k \in \{1, \ldots, K\}$, and the transmit signal $\boldsymbol{x}$ in order to maximize the extension of the achievable region for this kind of channels. A reasonable choice for $p(\boldsymbol{x}, \boldsymbol{u}_1, \cdots, \boldsymbol{u}_K)$ can be made that is based on a successive encoding scheme (cf. Section 2.2.2.2) and dirty paper coding in

[4] In [37] real-valued noise and interference were assumed. However, the result straightforwardly extends to circularly symmetric complex-valued noise and interference.

order to suppress interference caused by previously encoded users. This choice of statistics was first proposed by Caire et al. [23] for the case of single antenna receivers. Yu et al. [145] first characterized the set of rates that are achievable with this choice of statistics in a broadcast channel with an arbitrary number of antennas at the receivers. Here, a precise description is given of all variables involved in the encoding process, how these variables are related and all choices and assumptions that must be made on all signals in order to arrive at a joint probability function $p(\boldsymbol{x}, \boldsymbol{u}_1, \cdots, \boldsymbol{u}_K)$ that is based on dirty paper coding.

For the general successive encoding approach, the relation between the signals involved in the encoding process is illustrated in Fig. 2.7. As already pointed out in Section 2.2.2.2, successive encoding requires the introduction of variables $\boldsymbol{s}_k$, $k \in \{1, \ldots, K\}$, representing the signals actually sent by the users. As can be observed in Fig. 2.7, assuming that the encoding order is given by the user index, i.e., user 1 is encoded first and user K is the last encoded user, the joint probability density function of all these signals factorizes as follows,

$$\begin{aligned} &p(\boldsymbol{x}, \boldsymbol{s}_1, \cdots, \boldsymbol{s}_K, \boldsymbol{u}_1, \cdots, \boldsymbol{u}_K) = \\ &= p(\boldsymbol{x}|\boldsymbol{s}_1, \cdots, \boldsymbol{s}_K) \prod_{k=1}^{K} p(\boldsymbol{s}_k|\boldsymbol{u}_k, \boldsymbol{s}_1, \cdots, \boldsymbol{s}_{k-1}) p(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_K). \end{aligned} \tag{2.16}$$

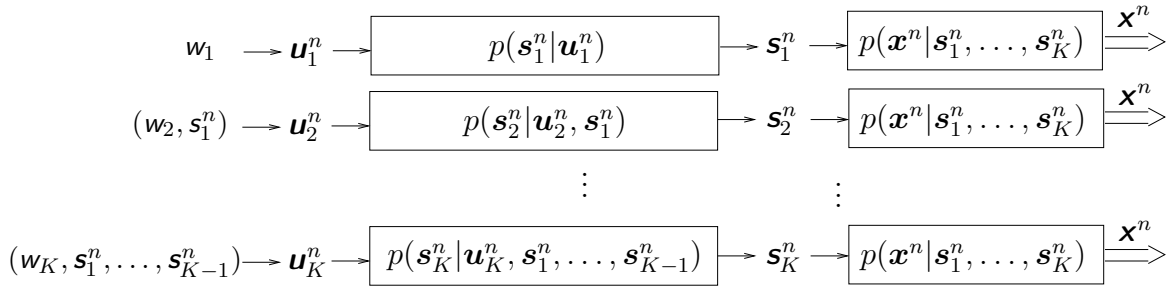

Figure 2.7: Successive coding for the Gaussian MIMO broadcast channel.

In order to simplify matters all signals can be chosen to be jointly Gaussian. It makes sense to chose $\boldsymbol{x}$ to be a deterministic function of the signals $\boldsymbol{s}_{1,\ldots,K}$ as otherwise information would be already lost before transmission. If the chosen function is linear, we can write $p(\boldsymbol{x}|\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K) = \delta(\boldsymbol{x} - \sum_{k=1}^{K} \boldsymbol{B}_k \boldsymbol{s}_k)$. Correspondingly, the signal received by user k' is given by

$$\boldsymbol{y}_{k'} = \boldsymbol{H}_{k'} \boldsymbol{B}_{k'} \boldsymbol{s}_{k'} + \boldsymbol{H}_{k'} \sum_{k<k'} \boldsymbol{B}_k \boldsymbol{s}_k + \boldsymbol{H}_{k'} \sum_{k>k'} \boldsymbol{B}_k \boldsymbol{s}_k + \boldsymbol{n}_{k'}.$$

Signals $\boldsymbol{s}_{1,\ldots,K}$ are chosen to be mutually uncorrelated with covariance matrix $\mathrm{E}\left\{\boldsymbol{s}_k \boldsymbol{s}_k^{\mathrm{H}}\right\} = \boldsymbol{I}_{m_k}$ and $m_k = \mathrm{Rank}\left\{\boldsymbol{H}_k\right\}$. Choosing these signals to be mutually uncorrelated is perfectly consistent with the dirty paper coding scheme, which delivers transmit signals that are uncorrelated with the known interference. Choosing these signals to be white does not represent any restriction as any wished correlation can be enforced by a convenient choice of the matrices $\boldsymbol{B}_{1,\ldots,K}$. Finally, choosing the dimension of these vectors to be equal to

the rank of the respective channel matrix is without loss of optimality as this constitutes the maximum number of dimensions over which information can be transmitted. As a consequence of these assumptions, the transmit power constraint can be written in terms of the beamforming matrices as

$$\sum_{k=1}^{K} \mathrm{Tr}\left\{\boldsymbol{B}_k \boldsymbol{B}_k^{\mathrm{H}}\right\} \leq P. \tag{2.17}$$

Once these choices have been made, all other statistical relations between signals follow from the successive application of the dirty paper coding scheme. For the first user, the received signal is given by

$$\boldsymbol{y}_1 = \boldsymbol{H}_1 \boldsymbol{B}_1 \boldsymbol{s}_1 + \boldsymbol{H}_1 \sum_{k>1} \boldsymbol{B}_k \boldsymbol{s}_k + \boldsymbol{n}_k.$$

We observe that the second and third terms on the right-hand side correspond to unknown interference and noise. Correspondingly, if $\boldsymbol{u}_1 = \boldsymbol{s}_1$, i.e., $p(\boldsymbol{s}_1|\boldsymbol{u}_1) = \delta(\boldsymbol{s}_1 - \boldsymbol{u}_1)$, is chosen, the following rate can be achieved that is optimum given a fixed beamforming matrix $\boldsymbol{B}_1$,

$$R_1 = \log_2 \left(\frac{|\boldsymbol{I}_{r_1} + \boldsymbol{H}_1 \sum_{k=1}^{K} \boldsymbol{\Sigma}_k \boldsymbol{H}_1^{\mathrm{H}}|}{|\boldsymbol{I}_{r_1} + \boldsymbol{H}_1 \sum_{k=2}^{K} \boldsymbol{\Sigma}_k \boldsymbol{H}_1^{\mathrm{H}}|} \right).$$

In the above expression $\boldsymbol{\Sigma}_k = \boldsymbol{B}_k \boldsymbol{B}_k^{\mathrm{H}}$ represents the transmit covariance matrix for user k. For notational convenience, the noise vector has been assumed to be white with unit-variance entries. Note that this assumption is without loss of generality as noise whitening is, for all practical purposes, an invertible operation that can always be reversed by the decoder. That is, from a broadcast channel with colored noise an equivalent broadcast channel with white noise can be obtained by whitening the noise at each receiver and regarding the products of each of the original channel matrices and the corresponding noise whitening filters as the channel matrices of the new "whitened" broadcast channel. Therefore, the white noise assumption shall hold by default for the rest of the discussion unless otherwise stated.

For the second user, the received signal is given by

$$\boldsymbol{y}_2 = \boldsymbol{H}_2 \boldsymbol{B}_2 \boldsymbol{s}_2 + \boldsymbol{H}_2 \boldsymbol{B}_1 \boldsymbol{s}_1 + \boldsymbol{H}_2 \sum_{k=3}^{K} \boldsymbol{B}_k \boldsymbol{s}_k + \boldsymbol{n}_k.$$

Here, the second term on the right-hand side represents the interference known at the transmitter. The third and fourth are the terms corresponding to unknown interference and noise. Now, we can identify these terms with the corresponding terms in Eq. 2.15, i.e.,

$$\boldsymbol{s} = \boldsymbol{H}_2 \boldsymbol{B}_2 \boldsymbol{s}_2, \quad \boldsymbol{u} = \boldsymbol{H}_2 \boldsymbol{B}_1 \boldsymbol{s}_1, \quad \boldsymbol{z} = \boldsymbol{H}_2 \sum_{k=3}^{K} \boldsymbol{B}_k \boldsymbol{s}_k + \boldsymbol{n}_k.$$

As pointed out in the previous section, in this case, optimality is achieved if the codebook is generated according to a random variable

$$\begin{aligned}\boldsymbol{v} &= \boldsymbol{s} + \boldsymbol{\Gamma u} \\ &= \boldsymbol{H}_2\boldsymbol{B}_2\boldsymbol{s}_2 + \boldsymbol{H}_2\boldsymbol{B}_2\boldsymbol{B}_2^{\mathrm{H}}\boldsymbol{H}_2^{\mathrm{H}}\left(\boldsymbol{I}_{r_2} + \boldsymbol{H}_2\sum_{k=2}^{K}\boldsymbol{B}_k\boldsymbol{B}_k^{\mathrm{H}}\boldsymbol{H}_2^{\mathrm{H}}\right)^{-1}\boldsymbol{H}_2\boldsymbol{B}_1\boldsymbol{s}_1.\end{aligned}$$

This random variable can be viewed as the actual codebook associated to the random variable $\boldsymbol{u}_2$ transformed by the product of the corresponding beamforming matrix and the channel, i.e., $\boldsymbol{v} = \boldsymbol{H}_2\boldsymbol{B}_2\boldsymbol{u}_2$. In fact, if $\boldsymbol{H}_2\boldsymbol{B}_2$ is invertible[5] we immediately conclude.

$$I(\boldsymbol{s}_2;\boldsymbol{y}_2|\boldsymbol{s}_1) = I(\boldsymbol{v};\boldsymbol{y}_2) - I(\boldsymbol{v};\boldsymbol{s}_1) = I(\boldsymbol{u}_2;\boldsymbol{y}_2) - I(\boldsymbol{u}_2;\boldsymbol{s}_1),$$

which confirms that by choosing

$$\boldsymbol{u}_2 = \boldsymbol{s}_2 + \boldsymbol{B}_2^{\mathrm{H}}\boldsymbol{H}_2^{\mathrm{H}}\left(\boldsymbol{I}_{r_2} + \boldsymbol{H}_2\sum_{k=2}^{K}\boldsymbol{B}_k\boldsymbol{B}_k^{\mathrm{H}}\boldsymbol{H}_2^{\mathrm{H}}\right)^{-1}\boldsymbol{H}_2\boldsymbol{B}_1\boldsymbol{s}_1, \tag{2.18}$$

the rate

$$R_2 = I(\boldsymbol{s}_2;\boldsymbol{y}_2|\boldsymbol{s}_1) = \log_2\left(\frac{|\boldsymbol{I}_{r_2} + \boldsymbol{H}_2\sum_{k=2}^{K}\boldsymbol{\Sigma}_k\boldsymbol{H}_2^{\mathrm{H}}|}{|\boldsymbol{I}_{r_2} + \boldsymbol{H}_2\sum_{k=3}^{K}\boldsymbol{\Sigma}_k\boldsymbol{H}_2^{\mathrm{H}}|}\right)$$

is achievable for user 2. Proceeding in the same way for the rest of successive encoding steps, the maximum achievable rate for user k' can be written as

$$R_{k'} = I(\boldsymbol{s}_{k'};\boldsymbol{y}_{k'}|\boldsymbol{s}_1,\cdots,\boldsymbol{s}_{k'-1}) = \log_2\left(\frac{|\boldsymbol{I}_{r_{k'}} + \boldsymbol{H}_{k'}\sum_{k=k'}^{K}\boldsymbol{\Sigma}_k\boldsymbol{H}_{k'}^{\mathrm{H}}|}{|\boldsymbol{I}_{r_{k'}} + \boldsymbol{H}_{k'}\sum_{k=k'+1}^{K}\boldsymbol{\Sigma}_k\boldsymbol{H}_{k'}^{\mathrm{H}}|}\right), \tag{2.19}$$

which can be attained by using a codebook generated according to the random variable

$$\boldsymbol{u}_{k'} = \boldsymbol{s}_{k'} + \boldsymbol{B}_{k'}^{\mathrm{H}}\boldsymbol{H}_{k'}^{\mathrm{H}}\left(\boldsymbol{I}_{r_{k'}} + \boldsymbol{H}_{k'}\sum_{k=k'}^{K}\boldsymbol{B}_k\boldsymbol{B}_k^{\mathrm{H}}\boldsymbol{H}_{k'}^{\mathrm{H}}\right)^{-1}\boldsymbol{H}_{k'}\sum_{k=1}^{k'-1}\boldsymbol{B}_k\boldsymbol{s}_k. \tag{2.20}$$

At this point we come back to Fig. 2.7 and recall Eq. 2.16. Due to the fact that $\boldsymbol{u}_{k'}$ is obtained by adding $\boldsymbol{s}_{k'}$ and linear transformations of $\boldsymbol{s}_{1,\ldots,k'-1}$, $\boldsymbol{s}_{k'}$ is a deterministic function of $\boldsymbol{u}_{k'}$ and $\boldsymbol{s}_{1,\ldots,k'-1}$, i.e., $\boldsymbol{s}_{k'} = f(\boldsymbol{u}_{k'},\boldsymbol{s}_1,\cdots,\boldsymbol{s}_{k'-1})$ and, therefore, $p(\boldsymbol{s}_{k'}|\boldsymbol{u}_{k'},\boldsymbol{s}_1,\cdots,\boldsymbol{s}_{k'-1}) = \delta(\boldsymbol{s}_{k'} - f(\boldsymbol{u}_{k'},\boldsymbol{s}_1,\cdots,\boldsymbol{s}_{k'-1}))$. In turn, the joint probability density function $p(\boldsymbol{u}_1,\cdots,\boldsymbol{u}_K)$ is completely characterized by Eq. 2.20 and the distribution initially assumed for the signals $\boldsymbol{s}_{1,\ldots,K}$. Thus, we observe that $p(\boldsymbol{x},\boldsymbol{u}_1,\cdots,\boldsymbol{u}_K)$ is the result of assuming a certain encoding order, Gaussian distribution for all variables, the choice of beamforming matrices, certain

[5]Due to the fact that the number of columns in $\boldsymbol{B}_2$ is equal to the rank of $\boldsymbol{H}_2$, $\boldsymbol{H}_2\boldsymbol{B}_2$ is always invertible unless the columns of $\boldsymbol{B}_2$ are linearly dependent or some of these columns lie in the nullspace of $\boldsymbol{H}_2$. In the first case the number of columns could be reduced in order to obtain a full-rank matrix $\boldsymbol{B}_2$ without loss in performance. The second case means that resources are wasted by transmitting signals over directions that will never reach the receiver. This does not make sense and should be avoided.

assumptions made on the statistics of the signals $\boldsymbol{s}_{1,\ldots,K}$ and application of the dirty paper coding scheme. For a particular choice of statistics the achievable region is given by

$$\mathcal{R}^{\text{Marton}} = \left\{ \boldsymbol{\rho} \in \mathbb{R}_+ : \sum_{i \in \mathcal{S}} R_i \leq \sum_{i \in \mathcal{S}} \left(I(\boldsymbol{u}_i; \boldsymbol{y}_i) - h(\boldsymbol{u}_i) \right) + h(\boldsymbol{u}_{i \in \mathcal{S}}), \forall \mathcal{S} \subseteq \{1, \ldots, K\} \right\}, \tag{2.21}$$

where $\boldsymbol{\rho} = [R_1, \cdots, R_K]^{\mathrm{T}}$, $h(\boldsymbol{u}_i)$ is the differential entropy of $\boldsymbol{u}_i$ and $h(\boldsymbol{u}_{i \in \mathcal{S}})$ represents the joint entropy of all variables with indexes in $\mathcal{S}$. This is a generalization of the Marton region for the two-user case discussed in Section 2.1.3.1. Observe that this region has the form of a convex polytope in a K-dimensional space since it is defined as the intersection of a number of half-spaces. The rates given by Eq. 2.19 represent only a vertex of this region. The other vertices are achieved by varying the encoding order while keeping $p(\boldsymbol{x}, \boldsymbol{u}_1, \cdots, \boldsymbol{u}_K)$ fixed. Points on the facets can also be achieved by performing simultaneous encoding (cf. Section 2.1.3.1). The following example shall illustrate some of these details.

Assume a two-user Gaussian broadcast channel with channel matrices

$$\boldsymbol{H}_1 = \begin{bmatrix} 1 & 1/2 & 2 \\ -1 & 2 & 1/4 \end{bmatrix}, \quad \boldsymbol{H}_2 = \begin{bmatrix} 1/2 & 3 & 1/3 \\ 1 & -1 & -2 \end{bmatrix}. \tag{2.22}$$

If we choose

$$\boldsymbol{B}_1 = \begin{bmatrix} 1 & -1 \\ 1 & 2 \\ 2 & 1/2 \end{bmatrix}, \quad \boldsymbol{B}_2 = \begin{bmatrix} -1 & 1/3 \\ 1 & -1/3 \\ -1/2 & -1 \end{bmatrix}$$

to be the respective beamforming matrices, the auxiliary signals $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ are chosen to be uncorrelated and white, as discussed above, and apply dirty paper coding under the assumption that user 1 is encoded first, the resulting statistics $p(\boldsymbol{x}, \boldsymbol{u}_1, \boldsymbol{u}_2)$ give rise to the Marton region whose boundary is depicted by the solid line in Fig. 2.8. There, the vertex surrounded by the circle is reached if for these statistics user 1 is actually coded first. In this case, the rates are given by

$$\begin{aligned} R_1 &= I(\boldsymbol{u}_1; \boldsymbol{y}_1) = \log_2 \left(\frac{|\boldsymbol{I}_2 + \boldsymbol{H}_1 \boldsymbol{\Sigma}_1 \boldsymbol{H}_1^{\mathrm{H}} + \boldsymbol{H}_1 \boldsymbol{\Sigma}_2 \boldsymbol{H}_1^{\mathrm{H}}|}{|\boldsymbol{I}_2 + \boldsymbol{H}_1 \boldsymbol{\Sigma}_2 \boldsymbol{H}_1^{\mathrm{H}}|} \right), \\ R_2 &= I(\boldsymbol{u}_2; \boldsymbol{y}_2) - I(\boldsymbol{u}_1; \boldsymbol{u}_2) = I(\boldsymbol{u}_2; \boldsymbol{y}_2 | \boldsymbol{u}_1) = \log_2 \left(|\boldsymbol{I}_2 + \boldsymbol{H}_2 \boldsymbol{\Sigma}_2 \boldsymbol{H}_2^{\mathrm{H}}| \right), \end{aligned}$$

which are those obtained from application of the dirty paper coding scheme. Alternatively, for these statistics, we could choose to code user 2 first (cf. Section). In that case we would obtain the vertex marked by the square. Now, the resulting rates are given by

$$\begin{aligned} R_1 &= I(\boldsymbol{u}_1; \boldsymbol{y}_1) - I(\boldsymbol{u}_1; \boldsymbol{u}_2) = \\ &= \log_2 \left(\frac{|\boldsymbol{I}_2 + \boldsymbol{H}_1 \boldsymbol{\Sigma}_1 \boldsymbol{H}_1^{\mathrm{H}} + \boldsymbol{H}_1 \boldsymbol{\Sigma}_2 \boldsymbol{H}_1^{\mathrm{H}}|}{|\boldsymbol{I}_2 + \boldsymbol{H}_1 \boldsymbol{\Sigma}_2 \boldsymbol{H}_1^{\mathrm{H}}|} \right) - \\ &\quad - \log_2 \left(|\boldsymbol{I}_2 + \boldsymbol{H}_2 \boldsymbol{\Sigma}_2 \boldsymbol{H}_2^{\mathrm{H}} \left(\boldsymbol{I}_2 + \boldsymbol{H}_2 \boldsymbol{\Sigma}_2 \boldsymbol{H}_2^{\mathrm{H}} \right)^{-1} \boldsymbol{H}_2 \boldsymbol{\Sigma}_1 \boldsymbol{H}_2^{\mathrm{H}} \left(\boldsymbol{I}_2 + \boldsymbol{H}_2 \boldsymbol{\Sigma}_2 \boldsymbol{H}_2^{\mathrm{H}} \right)^{-1} | \right), \\ R_2 &= I(\boldsymbol{u}_2; \boldsymbol{y}_2) = \log_2 \left(|\boldsymbol{I}_2 + \boldsymbol{H}_2 \boldsymbol{\Sigma}_2 \boldsymbol{H}_2^{\mathrm{H}} + \boldsymbol{H}_2 \boldsymbol{\Sigma}_2 \boldsymbol{H}_2^{\mathrm{H}} \left(\boldsymbol{I}_2 + \boldsymbol{H}_2 \boldsymbol{\Sigma}_2 \boldsymbol{H}_2^{\mathrm{H}} \right)^{-1} \boldsymbol{H}_2 \boldsymbol{\Sigma}_1 \boldsymbol{H}_2^{\mathrm{H}}| \right). \end{aligned}$$

Taking into account that $\boldsymbol{x} = \boldsymbol{B}_1\boldsymbol{s}_1 + \boldsymbol{B}_2\boldsymbol{s}_2$ and Eq. 2.18, we can write $\boldsymbol{x}$ in terms of $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ as

$$\boldsymbol{x} = \boldsymbol{B}_1\boldsymbol{u}_1 + \boldsymbol{B}_2\boldsymbol{u}_2 - \boldsymbol{B}_2\boldsymbol{B}_2^{\mathrm{H}}\boldsymbol{H}_2^{\mathrm{H}}\left(\boldsymbol{I}_2 + \boldsymbol{H}_2\boldsymbol{B}_2\boldsymbol{B}_2^{\mathrm{H}}\boldsymbol{H}_2^{\mathrm{H}}\right)^{-1}\boldsymbol{H}_2\boldsymbol{B}_1\boldsymbol{u}_1.$$

This expression holds for both encoding orders. If user 1 is encoded first, the signal intended for this user is $\boldsymbol{x}_1 = \boldsymbol{B}_1\boldsymbol{u}_1 = \boldsymbol{B}_1\boldsymbol{s}_1$, and that intended for user 2 is

$$\boldsymbol{x}_2 = \boldsymbol{B}_2\left(\boldsymbol{u}_2 - \boldsymbol{B}_2^{\mathrm{H}}\boldsymbol{H}_2^{\mathrm{H}}\left(\boldsymbol{I}_2 + \boldsymbol{H}_2\boldsymbol{B}_2\boldsymbol{B}_2^{\mathrm{H}}\boldsymbol{H}_2^{\mathrm{H}}\right)^{-1}\boldsymbol{H}_2\boldsymbol{B}_1\boldsymbol{u}_1\right) = \boldsymbol{B}_2\boldsymbol{s}_2.$$

By construction, these signals are uncorrelated. If user 2 is encoded first, the signal intended for user 1 is

$$\boldsymbol{x}_1 = \left(\boldsymbol{B}_1 - \boldsymbol{B}_2\boldsymbol{B}_2^{\mathrm{H}}\boldsymbol{H}_2^{\mathrm{H}}\left(\boldsymbol{I}_2 + \boldsymbol{H}_2\boldsymbol{B}_2\boldsymbol{B}_2^{\mathrm{H}}\boldsymbol{H}_2^{\mathrm{H}}\right)^{-1}\boldsymbol{H}_2\boldsymbol{B}_1\right)\boldsymbol{u}_1,$$

and that intended for user 2 is $\boldsymbol{x}_2 = \boldsymbol{B}_2\boldsymbol{u}_2$. These two signals are correlated.

If, now, the statistics of the transmit signals are determined by assuming that user 2 is encoded first, the corresponding region is given by the dashed line in Fig. 2.8. Again, once the statistics are established, we can choose to code in the order assumed for the choice of the statistics, which leads to the vertex marked by the circle, or modify the encoding order, which leads to the vertex marked by the square. Points between the vertices can be theoretically attained in both regions by performing simultaneous encoding with the respective statistics.

Coming back to the general Gaussian broadcast channel with K users, let us define a permutation function $\pi : \{1, \ldots, K\} \rightarrow \{1, \ldots, K\}$ defining the order that is considered in order to determine $p(\boldsymbol{x}, \boldsymbol{u}_1, \cdots, \boldsymbol{u}_K)$, i.e., user $\pi(1)$ is encoded in the first place, user $\pi(2)$ in the second place, and so on. Besides, let $\mathcal{R}^{\mathrm{Marton}}(\pi, \boldsymbol{\Sigma}_{1,\ldots,K})$ be the Marton region obtained from the statistics induced by the permutation π and beamforming matrices $\boldsymbol{B}_{1,\ldots,K}$ such that $\boldsymbol{B}_k\boldsymbol{B}_k^{\mathrm{H}} = \boldsymbol{\Sigma}_k$.[6] For a given transmit power limit P the dirty paper coding region can be now formally defined as

$$\mathcal{R}^{\mathrm{DPC}}(P) = Co \bigcup_{\pi, \boldsymbol{\Sigma}_{1,\ldots,K}} \mathcal{R}^{\mathrm{Marton}}\left(\pi, \boldsymbol{\Sigma}_{1,\ldots,K}\right), \tag{2.23}$$

where Co represents the convex hull operator and the transmit covariance matrices $\boldsymbol{\Sigma}_{1,\ldots,K}$ satisfy the transmit power constraint (cf. Eq. 2.17). Thus, the dirty paper region is given by the convex hull of the union of Marton regions corresponding to all possible statistics that can be defined by application of the dirty paper coding scheme under a given power constraint. Eq. 2.23 represents the conceptual link between the dirty paper region, which has been extensively discussed in the literature in recent years (e.g., [23, 144, 128]), and the Marton achievability region introduced in [80]. Even though this connection has been noted in most of these recent works (cf. [23, 144, 132, 122]), none of these publications makes it explicit. Actually, in the literature, the DPC region is not defined as the convex hull of unions of Marton regions associated with statistics obtained from a DPC-based

[6] The statistics depend on the transmit covariance matrices and not on the particular choice of beamforming matrices.

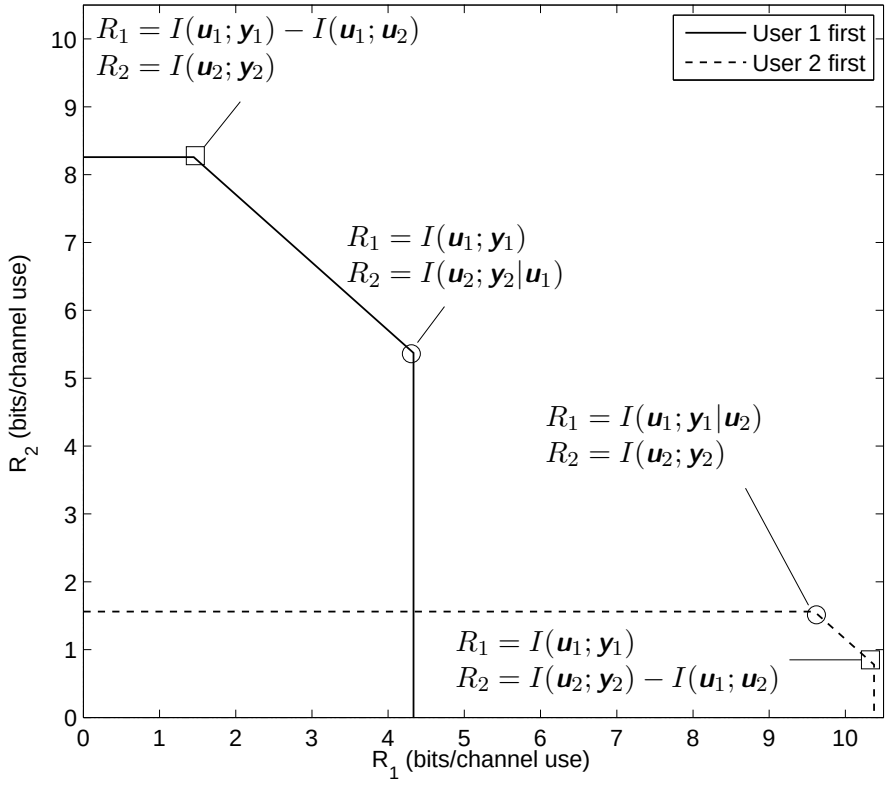

Figure 2.8: Marton regions for two different statistics of the transmit signals obtained by application of dirty paper coding with different orderings and equal beamforming matrices.

successive encoding scheme. Instead, it is defined as the convex hull of all rates that can be achieved by only considering DPC-based successive encoding as transmission scheme,

$$\mathcal{R}^{\mathrm{DPC}}(P) = \\ Co \bigcup_{\pi, \boldsymbol{\Sigma}_{1,\ldots,K}} \left\{ \boldsymbol{\rho} \in \mathbb{R}_+^K : R_{\pi(k')} \leq \log_2 \left(\frac{|\boldsymbol{I}_{r_{\pi(k')}} + \boldsymbol{H}_{\pi(k')} \sum_{k \geq k'} \boldsymbol{\Sigma}_{\pi(k)} \boldsymbol{H}_{\pi(k')}^{\mathrm{H}}|}{|\boldsymbol{I}_{r_{\pi(k')}} + \boldsymbol{H}_{\pi(k')} \sum_{k > k'} \boldsymbol{\Sigma}_{\pi(k)} \boldsymbol{H}_{\pi(k')}^{\mathrm{H}}|} \right) \right\}. \quad (2.24)$$

In other words, for a given ordering and choice of beamforming matrices only the vertex that corresponds to that ordering is considered in the definition of the dirty paper coding region. All other vertices are ignored. Thus, the dirty paper region as defined in the literature is included in the dirty paper region as defined by Eq. 2.23. Indeed, Eq. 2.23 considers all vertices of the Marton region associated with each choice of statistics and not just one. However, due to the fact that the dirty paper region as defined in Eq. 2.24 has been shown to be the actual capacity region, both definitions turn out to be equivalent. The only difference between both definitions is subtle but theoretically interesting. While the definition given by Eq. 2.24 implies the existence of points that can only be achieved by switching among different transmit strategies (time sharing), the definition given by

Eq. 2.23 suggests that, at least for some of these time-sharing points, coding schemes exist that render these points also achievable by just using a unique transmit strategy. This will be further discussed in the next section.

2.2.2.3 The dual multiple access channel

Recall the model for the broadcast channel given by Eq. 2.6 and assume that every user experiences a white noise with unit-variance components, i.e., $\mathrm{E}\{\boldsymbol{n}_k \boldsymbol{n}_k^{\mathrm{H}}\} = \boldsymbol{I}_{r_k}$, $\forall k$. For this broadcast channel, the dual multiple access channel is defined as

$$\boldsymbol{r} = \sum_{k=1}^{K} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{w}_k + \boldsymbol{z}, \tag{2.25}$$

where $\boldsymbol{r} \in \mathbb{C}^t$ is the vector of received signals, $\boldsymbol{w}_k \in \mathbb{C}^{r_k}$ denotes the vector of transmitted signals corresponding to user k and $\boldsymbol{z} \in \mathbb{C}^t$ is a realization of a random variable $\boldsymbol{z} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$ representing noise.

The channel just defined is a Gaussian multiple access channel. In general, a multiple access channel is a triple $(\mathcal{W}_1 \times \cdots \times \mathcal{W}_K, p(r|w_1, \cdots, w_k), \mathcal{R})$ formed by a cartesian product of input alphabets $\mathcal{W}_k$, $k \in \{1, \ldots, K\}$, a probability transition function $p(r|w_1, \cdots, w_k)$ and an output alphabet $\mathcal{R}$. In contrast to general broadcast channels, the capacity region of general multiple access channels is known [2]. For a fixed distribution of the inputs $p(w_1)p(w_2)\cdots p(w_K)$ the capacity region is given by

$$\begin{aligned} &\mathcal{R}^{\mathrm{MAC}}(p(w_1)p(w_2)\cdots p(w_K)) = \\ &\left\{ \boldsymbol{\rho} \in \mathbb{R}_+^K : \sum_{k \in \mathcal{S}} R_k \leq I(\mathsf{w}_{k\in\mathcal{S}}; \mathsf{r} | \mathsf{w}_{k\in\bar{\mathcal{S}}}), \forall \mathcal{S} \subseteq \{1, \ldots, K\} \right\}, \end{aligned} \tag{2.26}$$

where $\bar{\mathcal{S}} = \{1, \ldots, K\} \setminus \mathcal{S}$ and $I(\mathsf{w}_{k\in\mathcal{S}}; \mathsf{r} | \mathsf{w}_{k\in\bar{\mathcal{S}}})$ is the mutual information between the input variables with indexes in $\mathcal{S}$ and the output variable r conditioned on knowledge of the input variables with indexes in $\bar{\mathcal{S}}$. This region is a convex polytope in a K-dimensional space since it is defined as an intersection of half-spaces. Furthermore, it turns out that for each inequality in the definition there always exists, at least, one point in the region that achieves equality. Due to this property, this region is a polymatroid [121]. As we shall see in the next chapter, this polymatroidal structure turns out to be very useful for solving certain optimization problems defined on the elements of this region.[7] All points of the MAC region can be achieved by performing joint detection at the receiver. Alternatively, one can achieve just the vertices by performing successive decoding and rely on time sharing in order to achieve all other points on the facets and ridges of the region [40]. In order to illustrate this let us look at Fig. 2.9. This figure represents the capacity region of a two-user channel with fixed inputs. Vertex A can be achieved by first decoding information proceeding from user 1 and then, under knowledge of the signal received from this user, decoding the information sent by user 2. In turn, vertex B can be achieved by reversing the decoding order. That is,

[7] This property of the multiple access capacity region is not shared by the Marton region given by Eq. 2.21. For the Marton region there might be inequalities that are loose for all points in the region.

first, user 2 is decoded under the influence of the signal coming from user 1, which acts as interference. Then, information from user 1 is decoded by using the knowledge of the signal received from user 2. In order to achieve the points between vertex A and B the receiver must perform joint detection, i.e., it must search in the codebooks for a pair of transmit signals (w_1^n, w_2^n) that is jointly typical with the received sequence y^n. Alternatively, the receiver can perform successive decoding and switch between both decoding orders (time sharing) in order to reach any point on this segment.

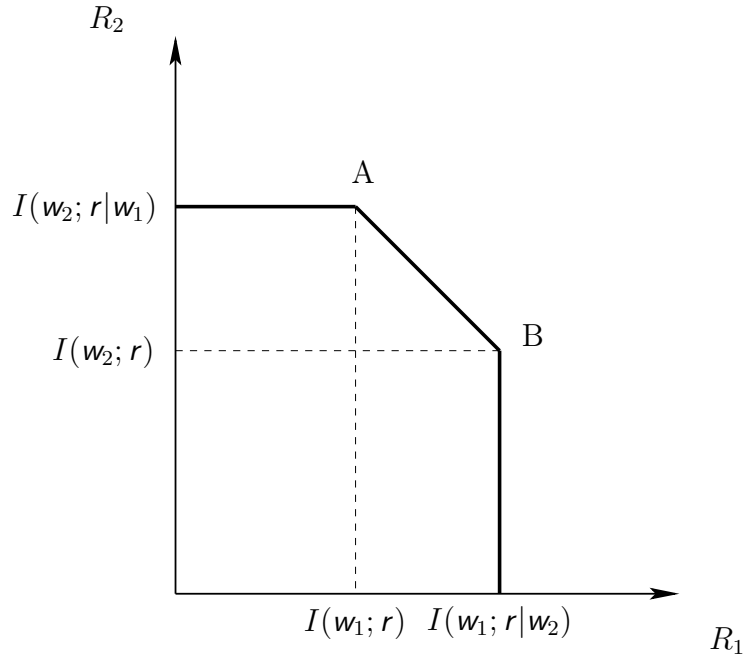

Figure 2.9: Capacity region of a two-user multiple access channel with fixed input distributions.

Hence, joint detection is not strictly necessary for attaining all points of the MAC capacity region. Elaborating on this idea an alternative definition of $\mathcal{R}^{\mathrm{MAC}}(p(w_1)p(w_2)\cdots p(w_K))$ can be given that is completely based on successive decoding and time sharing,

$$\begin{aligned}&\mathcal{R}^{\mathrm{MAC}}(p(w_1)p(w_2)\cdots p(w_K)) = \\ &Co\ \bigcup_{\bar{\pi}} \left\{ \boldsymbol{\rho} \in \mathbb{R}_+^K : R_{\bar{\pi}(k)} \leq I\left(\mathsf{w}_{\bar{\pi}(k)}; \mathsf{r} | \mathsf{w}_{\bar{\pi}(1)}, \ldots, \mathsf{w}_{\bar{\pi}(k-1)}\right),\ k = 1, \ldots, K \right\}. \end{aligned} \tag{2.27}$$

In this definition, $\bar{\pi} : \{1, \ldots, K\} \to \{1, \ldots, K\}$ is a permutation function that determines the order in which users are decoded, i.e., user $\bar{\pi}(1)$ is decoded first, user $\bar{\pi}(2)$ second, and so on. The union operation is over all possible permutations and the convex hull operation is needed in order to achieve those points for which a time-sharing strategy is required. Mathematically speaking, Eq. 2.26 defines a convex polytope in terms of supporting hyperplanes while Eq. 2.27 defines the same polytope as the convex hull of a given set of points, namely, the vertices.

For the Gaussian multiple access channel it is easily shown that optimum inputs are Gaussian distributed [40]. Thus, choosing the inputs in Eq. 2.25 to be zero-mean Gaussian

distributed and fixing the transmit covariance matrices of these inputs, Eq. 2.27 can be rewritten as

$$\mathcal{R}^{\mathrm{MAC}}\left(\boldsymbol{Q}_{1,\ldots,K}\right) = \tag{2.28}$$
$$Co \bigcup_{\bar{\pi}} \left\{ \boldsymbol{\rho} \in \mathbb{R}_+^K : R_{\bar{\pi}(k')} \leq \log_2 \left(\frac{|\boldsymbol{I}_t + \sum_{k \geq k'} \boldsymbol{H}_{\bar{\pi}(k)}^{\mathrm{H}} \boldsymbol{Q}_{\bar{\pi}(k)} \boldsymbol{H}_{\bar{\pi}(k)}|}{|\boldsymbol{I}_t + \sum_{k > k'} \boldsymbol{H}_{\bar{\pi}(k)}^{\mathrm{H}} \boldsymbol{Q}_{\bar{\pi}(k)} \boldsymbol{H}_{\bar{\pi}(k)}|} \right),\ k = 1, \ldots, K \right\},$$

where $\boldsymbol{Q}_k = \mathrm{E}\left\{\boldsymbol{w}_k \boldsymbol{w}_k^{\mathrm{H}}\right\}$ is the transmit covariance matrix of user k.

Once some of the basic properties of multiple access channels have been discussed, two theorems follow that establish the link between a Gaussian broadcast channel and its dual multiple access channel [128].

Theorem 2.2.1. *In the dual multiple access channel of a given broadcast channel, assume a decoding order $\bar{\pi}$ and a set of transmit covariance matrices $\boldsymbol{Q}_{1,\ldots,K}$ and let $\boldsymbol{\rho}^{\mathrm{MAC}}$ be the vector of achievable rates with this order and these matrices, i.e.,*

$$R_{\bar{\pi}(k')}^{\mathrm{MAC}} = \log_2 \left(\frac{|\boldsymbol{I}_t + \sum_{k \geq k'} \boldsymbol{H}_{\bar{\pi}(k)}^{\mathrm{H}} \boldsymbol{Q}_{\bar{\pi}(k)} \boldsymbol{H}_{\bar{\pi}(k)}|}{|\boldsymbol{I}_t + \sum_{k > k'} \boldsymbol{H}_{\bar{\pi}(k)}^{\mathrm{H}} \boldsymbol{Q}_{\bar{\pi}(k)} \boldsymbol{H}_{\bar{\pi}(k)}|} \right),\ \forall k' \in \{1, \ldots, K\}.$$

Then, there exists a set of transmit covariance matrices $\boldsymbol{\Sigma}_{1,\ldots,K}$ in the original broadcast channel such that, for the encoding order $\pi(k) = \bar{\pi}(K - k + 1)$, $\boldsymbol{\rho}^{\mathrm{DPC}} = \boldsymbol{\rho}^{\mathrm{MAC}}$, where

$$R_{\pi(k')}^{\mathrm{DPC}} = \log_2 \left(\frac{|\boldsymbol{I}_{r_{\pi(k')}} + \boldsymbol{H}_{\pi(k')} \sum_{k \geq k'} \boldsymbol{\Sigma}_{\pi(k)} \boldsymbol{H}_{\pi(k')}^{\mathrm{H}}|}{|\boldsymbol{I}_{r_{\pi(k')}} + \boldsymbol{H}_{\pi(k')} \sum_{k > k'} \boldsymbol{\Sigma}_{\pi(k)} \boldsymbol{H}_{\pi(k')}^{\mathrm{H}}|} \right),\ \forall k' \in \{1, \ldots, K\}.$$

Furthermore, $\mathrm{Tr}\left\{\sum_{k=1}^K \boldsymbol{\Sigma}_k\right\} \leq \mathrm{Tr}\left\{\sum_{k=1}^K \boldsymbol{Q}_k\right\}$.

According to this theorem, for a given constraint on the overall transmitted power, every rate vector that is achievable in the dual MAC by performing successive detection is also achievable in the original BC by performing successive encoding based on the dirty paper coding scheme. This is possible by setting the encoding order in the BC to be the reversed of the decoding order in the MAC. The next theorem states the converse, namely, if a certain rate vector can be achieved in the BC by applying DPC, that same vector can also be achieved in the dual MAC by performing successive detection with reversed ordering.

Theorem 2.2.2. *Given a Gaussian broadcast channel, assume an encoding order π and a set of transmit covariance matrices $\boldsymbol{\Sigma}_{1,\ldots,K}$ and let $\boldsymbol{\rho}^{\mathrm{DPC}}$ be the vector of rates that can be achieved with this order and these matrices using dirty paper coding, i.e.,*

$$R_{\pi(k')}^{\mathrm{DPC}} = \log_2 \left(\frac{|\boldsymbol{I}_{r_{\pi(k')}} + \boldsymbol{H}_{\pi(k')} \sum_{k \geq k'} \boldsymbol{\Sigma}_{\pi(k)} \boldsymbol{H}_{\pi(k')}^{\mathrm{H}}|}{|\boldsymbol{I}_{r_{\pi(k')}} + \boldsymbol{H}_{\pi(k')} \sum_{k > k'} \boldsymbol{\Sigma}_{\pi(k)} \boldsymbol{H}_{\pi(k')}^{\mathrm{H}}|} \right),\ \forall k' \in \{1, \ldots, K\}.$$

Then, there exists a set of transmit covariance matrices $\boldsymbol{Q}_{1,\ldots,K}$ in the dual multiple access channel such that, for the decoding order $\bar{\pi}(k) = \pi(K - k + 1)$, $\boldsymbol{\rho}^{\mathrm{MAC}} = \boldsymbol{\rho}^{\mathrm{DPC}}$, where

$$R_{\bar{\pi}(k')}^{\mathrm{MAC}} = \log_2 \left(\frac{|\boldsymbol{I}_t + \sum_{k \geq k'} \boldsymbol{H}_{\bar{\pi}(k)}^{\mathrm{H}} \boldsymbol{Q}_{\bar{\pi}(k)} \boldsymbol{H}_{\bar{\pi}(k)}|}{|\boldsymbol{I}_t + \sum_{k > k'} \boldsymbol{H}_{\bar{\pi}(k)}^{\mathrm{H}} \boldsymbol{Q}_{\bar{\pi}(k)} \boldsymbol{H}_{\bar{\pi}(k)}|} \right),\ \forall k' \in \{1, \ldots, K\}.$$

Furthermore, $\mathrm{Tr}\left\{\sum_{k=1}^{K}\boldsymbol{Q}_k\right\} \leq \mathrm{Tr}\left\{\sum_{k=1}^{K}\boldsymbol{\Sigma}_k\right\}$.

In [128] an specific algorithm is given in order to transform transmit covariance matrices achieving a certain rate vector in the dual MAC into transmit covariance matrices achieving that vector in the broadcast channel and vice versa. As a consequence of the first theorem we can write

$$\mathcal{R}^{\mathrm{MAC}}(P) = \bigcup_{\boldsymbol{Q}_{1,\ldots,K}} \mathcal{R}^{\mathrm{MAC}}\left(\boldsymbol{Q}_{1,\ldots,K}\right) \subseteq \mathcal{R}^{\mathrm{DPC}}(P) \quad \text{with} \quad \mathrm{Tr}\left\{\sum_{k=1}^{K}\boldsymbol{Q}_k\right\} \leq P. \tag{2.29}$$

By contrast, even if we use Eq. 2.24 as the definition of the DPC region, Theorem 2.2.2 alone does not imply the inclusion in the inverse direction. This is due to the convex hull operator in Eq. 2.24, which might incorporate rate vectors that lie outside $\mathcal{R}^{\mathrm{MAC}}(P)$. However, this possibility can be discarded by proving that $\mathcal{R}^{\mathrm{MAC}}(P)$ is a convex set. To this end, assume that $\boldsymbol{\rho}$ and $\bar{\boldsymbol{\rho}}$ are two rate vectors that belong to $\mathcal{R}^{\mathrm{MAC}}(P)$. Making use of Eq. 2.26, this implies that there are matrices $\boldsymbol{Q}_{1,\ldots,K}$ and $\bar{\boldsymbol{Q}}_{1,\ldots,K}$ such that

$$\sum_{k\in\mathcal{S}} R_k \leq \log_2\left(\left|\boldsymbol{I}_t + \sum_{k\in\mathcal{S}}\boldsymbol{H}_k^{\mathrm{H}}\boldsymbol{Q}_k\boldsymbol{H}_k\right|\right), \quad \sum_{k\in\mathcal{S}} \bar{R}_k \leq \log_2\left(\left|\boldsymbol{I}_t + \sum_{k\in\mathcal{S}}\boldsymbol{H}_k^{\mathrm{H}}\bar{\boldsymbol{Q}}_k\boldsymbol{H}_k\right|\right),$$

$\forall\mathcal{S}\subseteq\{1,\ldots,K\}$. Now, consider the rate vector $\hat{\boldsymbol{\rho}} = \mu\boldsymbol{\rho} + (1-\mu)\bar{\boldsymbol{\rho}}$ with $0\leq\mu\leq 1$. Due to concavity of $\log_2|\bullet|$ (cf. [40, Theorem 17.9.1]) we can apply Jensen's inequality to obtain,

$$\begin{aligned}
\sum_{k\in\mathcal{S}}\hat{R}_k &\leq \mu\log_2\left(\left|\boldsymbol{I}_t + \sum_{k\in\mathcal{S}}\boldsymbol{H}_k^{\mathrm{H}}\boldsymbol{Q}_k\boldsymbol{H}_k\right|\right) + (1-\mu)\log_2\left(\left|\boldsymbol{I}_t + \sum_{k\in\mathcal{S}}\boldsymbol{H}_k^{\mathrm{H}}\bar{\boldsymbol{Q}}_k\boldsymbol{H}_k\right|\right) \\
&\leq \log_2\left(\left|\boldsymbol{I}_t + \sum_{k\in\mathcal{S}}\boldsymbol{H}_k^{\mathrm{H}}\left(\mu\boldsymbol{Q}_k + (1-\mu)\bar{\boldsymbol{Q}}_k\right)\boldsymbol{H}_k\right|\right) \\
&= \log_2\left(\left|\boldsymbol{I}_t + \sum_{k\in\mathcal{S}}\boldsymbol{H}_k^{\mathrm{H}}\hat{\boldsymbol{Q}}_k\boldsymbol{H}_k\right|\right).
\end{aligned}$$

Since $\mathrm{Tr}\left\{\sum_{k=1}^{K}\hat{\boldsymbol{Q}}_k\right\} = \mu\mathrm{Tr}\left\{\sum_{k=1}^{K}\boldsymbol{Q}_k\right\} + (1-\mu)\mathrm{Tr}\left\{\sum_{k=1}^{K}\bar{\boldsymbol{Q}}_k\right\} \leq P$, we conclude that $\hat{\boldsymbol{\rho}}$ is also in $\mathcal{R}^{\mathrm{MAC}}(P)$, from which the convexity of $\mathcal{R}^{\mathrm{MAC}}(P)$ follows. As a consequence of this property and Theorem 2.2.2 we can write $\mathcal{R}^{\mathrm{DPC}}(P)\subseteq\mathcal{R}^{\mathrm{MAC}}(P)$ and, therefore, $\mathcal{R}^{\mathrm{DPC}}(P) = \mathcal{R}^{\mathrm{MAC}}(P)$, at least for the definition of $\mathcal{R}^{\mathrm{DPC}}(P)$ given by Eq. 2.24. Note that this might not hold if $\mathcal{R}^{\mathrm{DPC}}(P)$ is given by Eq. 2.23, as this definition incorporates points that, are not attained by direct application of the dirty paper coding scheme. Only as a result of the fact that $\mathcal{R}^{\mathrm{DPC}}(P)$ as defined in Eq. 2.24 is the actual capacity region of the BC, we can state that both definitions are equivalent.

As it has been already mentioned, fixing the transmit covariance matrices in the multiple access channel a polymatroidal region $\mathcal{R}^{\mathrm{MAC}}\left(\boldsymbol{Q}_{1,\ldots,K}\right)$ is obtained. In addition, due to Theorem 2.2.1 the vertices of this polymatroid can also be achieved in the broadcast channel. More precisely, given the vertex in the MAC region corresponding to a decoding

order $\bar{\pi}$, that vertex can be achieved in the BC by computing the transmit covariance matrices $\boldsymbol{\Sigma}_{1,\ldots,K}$ according to the transformations given in [128] and successively encoding the users according to a DPC scheme with reversed order $\bar{\pi}$. While in the MAC all vertices are achieved with the same statistics, in the BC, every vertex requires a different set of transmit covariance matrices. Furthermore, if transmission is based on successive DPC coding and only the statistics associated with the vertices are available, points between the vertices are only achievable by performing time sharing. Alternatively, given the statistics of each of the vertices, simultaneous encoding or general successive encoding might be considered. Before concluding this section, let us take a closer look at the relationship between $\mathcal{R}^{\text{MAC}}\left(\boldsymbol{Q}_{1,\ldots,K}\right)$ and the Marton regions $\mathcal{R}^{\text{Marton}}\left(\pi, \boldsymbol{\Sigma}_{1,\ldots,K}\right)$ that are obtained from the statistics of each of the vertices in the broadcast channel. To this end, we shall first show some numerical examples and we will finish by stating an interesting conjecture that we will be able to prove for a simple case. Consider the broadcast channel with channel matrices given by

$$\boldsymbol{H}_1 = \begin{bmatrix} 0.7812 & -1.1878 \\ 0.5690 & -2.2023 \end{bmatrix}, \quad \boldsymbol{H}_2 = \begin{bmatrix} -0.8217 & 0.9863 \\ -0.2656 & -0.5186 \end{bmatrix}.$$

In Fig. 2.10 regions are plotted for two different choices of statistics in the dual MAC. The regions labeled as "Example 1" correspond to the following choice of covariance matrices,

$$\boldsymbol{Q}_1^{\text{Ex1}} = \begin{bmatrix} 0.2276 & 0.1165 \\ 0.1165 & 2.2472 \end{bmatrix}, \quad \boldsymbol{Q}_2^{\text{Ex1}} = \begin{bmatrix} 4.8710 & 3.3979 \\ 3.3979 & 2.6542 \end{bmatrix}.$$

Performing successive detection with ordering $\bar{\pi}(1) = 1$, $\bar{\pi}(2) = 2$ the point A_1 is achieved. This point is achieved in the BC by successively encoding users in the order $\pi(1) = 2$, $\pi(2) = 1$ according to the DPC scheme with covariance matrices

$$\boldsymbol{\Sigma}_1^{\text{A}_1} = \begin{bmatrix} 0.5510 & 0.8851 \\ 0.8851 & 1.5270 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2^{\text{A}_1} = \begin{bmatrix} 7.2116 & -1.4692 \\ -1.4692 & 0.7104 \end{bmatrix}.$$

Keeping these statistics fixed all points of the Marton region delimited by the solid line passing through A_1 can be achieved by using simultaneous coding or successive encoding in reversed order, i.e., no time sharing is needed. If the order $\bar{\pi}(1) = 2$, $\bar{\pi}(2) = 1$ is considered in the MAC, point B_1 is attained. The same point is achieved in the BC with DPC, reversed ordering and transmit matrices

$$\boldsymbol{\Sigma}_1^{\text{B}_1} = \begin{bmatrix} 0.6796 & -0.4894 \\ -0.4894 & 2.5536 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2^{\text{B}_1} = \begin{bmatrix} 6.4447 & 1.4068 \\ 1.4068 & 0.3220 \end{bmatrix}.$$

The Marton region corresponding to these statistics is delimited by the dotted line passing through point B_1. Note that if we limit ourselves to use a successive DPC scheme in the BC, only points A_1 and B_1 are directly achievable on the sum-rate segment. For all other points time sharing is required. However, considering all points in the Marton region associated to the statistics of any vertex, at least, some fraction of this segment can be achieved directly or, as this example shows, it might even occur that the DPC region is actually enlarged.

In "Example 1" the sum of the traces of the transmit covariance matrices amount to 10. In "Example 2" the covariance matrices are considered that maximize the sum rate for this power, i.e.,

$$\left\{\boldsymbol{Q}_1^{\mathrm{Ex2}}, \boldsymbol{Q}_2^{\mathrm{Ex2}}\right\} = \max_{\{\boldsymbol{Q}_1, \boldsymbol{Q}_2\}} \log_2\left(\left|\boldsymbol{I}_2 + \boldsymbol{H}_1^{\mathrm{H}}\boldsymbol{Q}_1\boldsymbol{H}_1 + \boldsymbol{H}_2^{\mathrm{H}}\boldsymbol{Q}_2\boldsymbol{H}_2\right|\right),$$

subject to $\mathrm{Tr}\left\{\boldsymbol{Q}_1 + \boldsymbol{Q}_2\right\} \leq 10$. Solving this convex optimization problem we obtain

$$\boldsymbol{Q}_1^{\mathrm{Ex2}} = \begin{bmatrix} 1.1655 & 2.4087 \\ 2.4087 & 4.9781 \end{bmatrix}, \quad \boldsymbol{Q}_2^{\mathrm{Ex2}} = \begin{bmatrix} 2.2671 & 1.8981 \\ 1.8981 & 1.5894 \end{bmatrix}.$$

Now, computation of the transmit covariance matrices for the vertices A_2 and B_2 in the broadcast channel yields

$$\boldsymbol{\Sigma}_1^{\mathrm{A}_2} = \begin{bmatrix} 0.2834 & 1.1744 \\ 1.1744 & 4.8661 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2^{\mathrm{A}_2} = \begin{bmatrix} 3.7903 & -2.0040 \\ -2.0040 & 1.0602 \end{bmatrix},$$

and

$$\boldsymbol{\Sigma}_1^{\mathrm{B}_2} = \begin{bmatrix} 0.6498 & -1.9055 \\ -1.9055 & 5.5883 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2^{\mathrm{B}_2} = \begin{bmatrix} 3.4239 & 1.0758 \\ 1.0758 & 0.3380 \end{bmatrix}.$$

Surprisingly, as we see in Fig. 2.10, the Marton regions corresponding to these two different statistics are equal and exactly as large as the MAC region. This suggests the following conjecture.

Conjecture 2.2.3. *Given a Gaussian broadcast channel, let $\boldsymbol{Q}_{1,\ldots,K}$ be the set of covariance matrices that maximize the sum rate in the dual MAC for a given transmit power constraint and let $\boldsymbol{\Sigma}_{1,\ldots,K}$ be a set of transmit covariance matrices achieving any of the vertices of the polymatroid $\mathcal{R}^{\mathrm{MAC}}\left(\boldsymbol{Q}_{1,\ldots,K}\right)$ in the BC by encoding users successively with order π according to the DPC scheme. Then,*

$$\mathcal{R}^{\mathrm{Marton}}\left(\pi, \boldsymbol{\Sigma}_{1,\ldots,K}\right) = \mathcal{R}^{\mathrm{MAC}}\left(\boldsymbol{Q}_{1,\ldots,K}\right).$$

For reasons that will become apparent in the next chapter, some of the time-sharing regions on the boundary of $\mathcal{R}^{\mathrm{DPC}}(P)$ as defined in Eq. 2.24 coincide with sum-rate maximizing statistics for all users or groups of users. Thus, the most significant consequence of Conjecture 2.2.3, should it be true, is that all points on these regions can be achieved directly. A similar result will be presented for the rest of time-sharing points of $\mathcal{R}^{\mathrm{DPC}}(P)$ in the next chapter, whose validity will be shown for a tractable setting. As a consequence, we conjecture that no time-sharing is needed in order to achieve any point in $\mathcal{R}^{\mathrm{DPC}}(P)$ or, in other words, the operator Co can be removed from the definition of $\mathcal{R}^{\mathrm{DPC}}(P)$ if Eq. 2.23 is used instead of Eq. 2.24.

In the following we demonstrate the validity of Conjecture 2.2.3 for broadcast channels with two users and single receive antennas. In this proof we make use of the duality transformations between MAC and BC and of the matrix inversion lemma. Both results are summarized in Appendix A.1. Let us consider the BC given by

$$\begin{aligned} y_1 &= \boldsymbol{h}_1^{\mathrm{H}}\boldsymbol{x} + n_1, \\ y_2 &= \boldsymbol{h}_2^{\mathrm{H}}\boldsymbol{x} + n_2, \end{aligned}$$

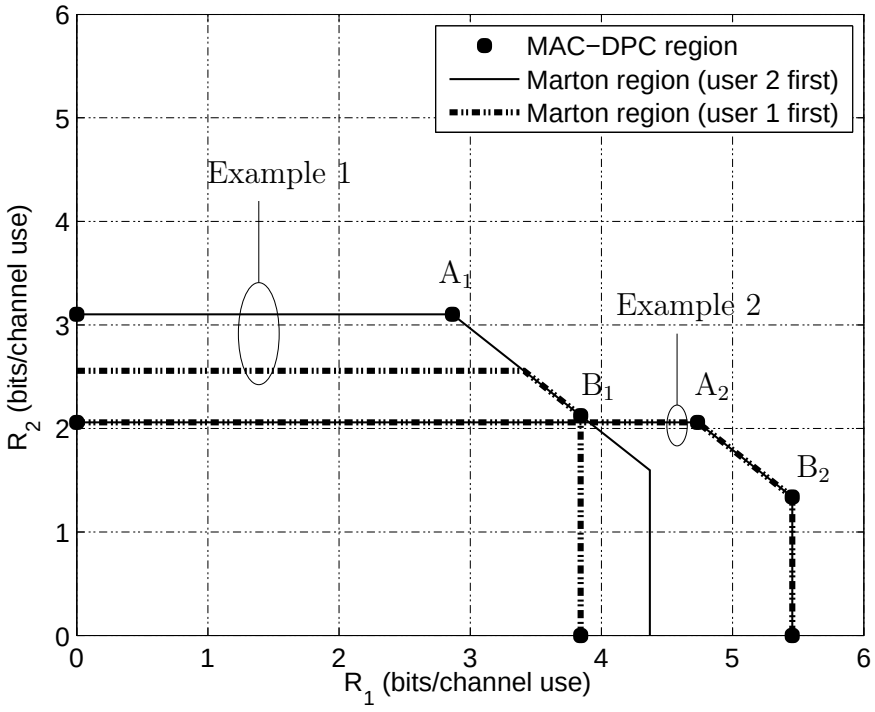

Figure 2.10: MAC capacity regions and associated Marton regions.

with unit-variance noises processes n_1 and n_2. The dual MAC for this channel reads

$$\boldsymbol{r} = \boldsymbol{h}_1 w_1 + \boldsymbol{h}_2 w_2 + z.$$

For a given power constraint P let q_1 and q_2 be maximizers of

$$\log_2 \left(|\boldsymbol{I}_t + q_1 \boldsymbol{h}_1 \boldsymbol{h}_1^{\mathrm{H}} + q_2 \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}}| \right).$$

Solving the Karush-Kuhn-Tucker (KKT) conditions [13] for the sum-rate optimization problem we obtain,

$$\boldsymbol{h}_1^{\mathrm{H}} \left(\boldsymbol{I}_t + q_1 \boldsymbol{h}_1 \boldsymbol{h}_1^{\mathrm{H}} + q_2 \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}} \right)^{-1} \boldsymbol{h}_1 = \boldsymbol{h}_2^{\mathrm{H}} \left(\boldsymbol{I}_t + q_1 \boldsymbol{h}_1 \boldsymbol{h}_1^{\mathrm{H}} + q_2 \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}} \right)^{-1} \boldsymbol{h}_2, \tag{2.30}$$

as a necessary condition for the optimality of q_1 and q_2 provided that both are greater than zero.[8] Now, applying the matrix inversion lemma [58] to both sides of the equation we get

$$\frac{\boldsymbol{h}_1^{\mathrm{H}} \left(\boldsymbol{I}_t + q_2 \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}} \right)^{-1} \boldsymbol{h}_1}{1 + q_1 \boldsymbol{h}_1^{\mathrm{H}} \left(\boldsymbol{I}_t + q_2 \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}} \right)^{-1} \boldsymbol{h}_1} = \frac{\boldsymbol{h}_2^{\mathrm{H}} \left(\boldsymbol{I}_t + q_1 \boldsymbol{h}_1 \boldsymbol{h}_1^{\mathrm{H}} \right)^{-1} \boldsymbol{h}_2}{1 + q_2 \boldsymbol{h}_2^{\mathrm{H}} \left(\boldsymbol{I}_t + q_1 \boldsymbol{h}_1 \boldsymbol{h}_1^{\mathrm{H}} \right)^{-1} \boldsymbol{h}_2}.$$

This expression can be further simplified by applying the matrix inversion lemma again on $(\boldsymbol{I}_t + q_2 \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}})^{-1}$ and $(\boldsymbol{I}_t + q_1 \boldsymbol{h}_1 \boldsymbol{h}_1^{\mathrm{H}})^{-1}$. Doing so, after some algebra we finally obtain

$$\|\boldsymbol{h}_1\|_2^2 \left(1 + q_2 \|\boldsymbol{h}_2\|_2^2 \right) - q_2 |\boldsymbol{h}_1^{\mathrm{H}} \boldsymbol{h}_2|^2 = \|\boldsymbol{h}_2\|_2^2 \left(1 + q_1 \|\boldsymbol{h}_1\|_2^2 \right) - q_1 |\boldsymbol{h}_1^{\mathrm{H}} \boldsymbol{h}_2|^2. \tag{2.31}$$

[8] If one of the powers is zero the polymatroid collapses to a segment on one of the axes. This case is trivial.

Assume that in the MAC successive detection is performed and that user 1 is detected first. The corresponding rate vector would be point A in Fig. 2.9. In particular, the rate achieved by user 1 at this point is given by

$$R_1^{\mathrm{A}} = \log_2\left(|\boldsymbol{I}_t + q_1\boldsymbol{h}_1\boldsymbol{h}_1^{\mathrm{H}} + q_2\boldsymbol{h}_2\boldsymbol{h}_2^{\mathrm{H}}|\right) - \log_2\left(|\boldsymbol{I}_t + q_2\boldsymbol{h}_2\boldsymbol{h}_2^{\mathrm{H}}|\right). \tag{2.32}$$

Considering the duality transformations given in [128] and summarized in Appendix A.1, the transmit covariance matrices that would achieve this point in the BC provided that DPC is used for transmission can be written as

$$\boldsymbol{\Sigma}_1 = q1\frac{\left(\boldsymbol{I}_t + q_2\boldsymbol{h}_2\boldsymbol{h}_2^{\mathrm{H}}\right)^{-1}\boldsymbol{h}_1\boldsymbol{h}_1^{\mathrm{H}}\left(\boldsymbol{I}_t + q_2\boldsymbol{h}_2\boldsymbol{h}_2^{\mathrm{H}}\right)^{-1}}{\boldsymbol{h}_1^{\mathrm{H}}\left(\boldsymbol{I}_t + q_2\boldsymbol{h}_2\boldsymbol{h}_2^{\mathrm{H}}\right)^{-1}\boldsymbol{h}_1}, \quad \boldsymbol{\Sigma}_2 = q_2\frac{\boldsymbol{h}_2\boldsymbol{h}_2^{\mathrm{H}}}{\|\boldsymbol{h}_2\|_2^2}(1 + \boldsymbol{h}_2^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_2). \tag{2.33}$$

Recalling the discussion in Section 2.2.2.2 and considering that at point A user 1 is the last encoded user we can write

$$R_1^{\mathrm{A}} = I(u_1; y_1) - I(u_1; u_2) = I(s_1; y_1) = \log_2(1 + \boldsymbol{h}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_1) \tag{2.34}$$

where $u_2 = s_2$ and

$$u_1 = s_1 + \boldsymbol{b}_1^{\mathrm{H}}\boldsymbol{h}_1(1 + \boldsymbol{h}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_1)^{-1}\boldsymbol{h}_1^{\mathrm{H}}\boldsymbol{b}_2 s_2.$$

There, $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are two transmit beamformers such that $\boldsymbol{\Sigma}_1 = \boldsymbol{b}_1\boldsymbol{b}_1^{\mathrm{H}}$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{b}_2\boldsymbol{b}_2^{\mathrm{H}}$, and s_1 and s_2 are two uncorrelated unit-variance random variables. Coming back to the dual MAC, point B is achieved by detecting user 2 first, the rate achieved by user 1 at this point is given by

$$R_1^{\mathrm{B}} = \log_2\left(|\boldsymbol{I}_t + q_1\boldsymbol{h}_1\boldsymbol{h}_1^{\mathrm{H}}|\right). \tag{2.35}$$

Now, consider the point in the Marton region $\mathcal{R}^{\mathrm{Marton}}(\pi, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$, with statistics determined by the ordering $\pi(1) = 2$, $\pi(2) = 1$, that is achieved by encoding user 1 in the first place, i.e., choosing the inverse ordering for the actual encoding of users. The rate for user 1 at this point is given by $\bar{R}_1^{\mathrm{B}} = I(u_1; y_1)$. We want to show that $\bar{R}_1^{\mathrm{B}} = R_1^{\mathrm{B}}$ or, equivalently, $R_1^{\mathrm{B}} - R_1^{\mathrm{A}} = I(u_1; u_2)$. $R_1^{\mathrm{B}} - R_1^{\mathrm{A}}$ is easily computed from Eq. 2.32 and Eq. 2.35 and

$$I(u_1; u_2) = \log_2\left(1 + \frac{\boldsymbol{h}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_1\boldsymbol{h}_1^{\mathrm{H}}\boldsymbol{\Sigma}_2\boldsymbol{h}_1}{\left(1 + \boldsymbol{h}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_1\right)^2}\right).$$

Thus, we must show that

$$\begin{aligned}\log_2\left(|\boldsymbol{I}_t + q_1\boldsymbol{h}_1\boldsymbol{h}_1^{\mathrm{H}}||\boldsymbol{I}_t + q_2\boldsymbol{h}_2\boldsymbol{h}_2^{\mathrm{H}}||\boldsymbol{I}_t + q_1\boldsymbol{h}_1\boldsymbol{h}_1^{\mathrm{H}} + q_2\boldsymbol{h}_2\boldsymbol{h}_2^{\mathrm{H}}|^{-1}\right) = \\ = \log_2\left(1 + \frac{\boldsymbol{h}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_1\boldsymbol{h}_1^{\mathrm{H}}\boldsymbol{\Sigma}_2\boldsymbol{h}_1}{\left(1 + \boldsymbol{h}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_1\right)^2}\right)\end{aligned} \tag{2.36}$$

holds under the condition given by Eq. 2.31. Removing logarithms and multiplying by $(1 + \boldsymbol{h}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_1)^2$ on both sides of the equality we get

$$|\boldsymbol{I}_t + q_1\boldsymbol{h}_1\boldsymbol{h}_1^{\mathrm{H}}|\left(1 + \boldsymbol{h}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_1\right) = \left(1 + \boldsymbol{h}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_1\right)^2 + \boldsymbol{h}_1^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_1\boldsymbol{h}_1^{\mathrm{H}}\boldsymbol{\Sigma}_2\boldsymbol{h}_1, \tag{2.37}$$

where we have used the fact that the rates given by Eq. 2.32 and Eq. 2.34 are equal and, therefore,

$$1 + \boldsymbol{h}_1^{\mathrm{H}} \boldsymbol{\Sigma}_1 \boldsymbol{h}_1 = |\boldsymbol{I}_t + q_2 \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}}|^{-1} |\boldsymbol{I}_t + q_1 \boldsymbol{h}_1 \boldsymbol{h}_1^{\mathrm{H}} + q_2 \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}}|.$$

Now, substituting Eqs. 2.33 in Eq. 2.37 and after some algebra we obtain

$$\begin{aligned}
&\|\boldsymbol{h}_2\|_2^2 \left(1 + q_1 \|\boldsymbol{h}_1\|_2^2\right) \left(1 + q_1 \boldsymbol{h}_1^{\mathrm{H}} \left(\boldsymbol{I}_t + q_2 \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}}\right)^{-1} \boldsymbol{h}_1\right) = \\
&q_1 q_2 |\boldsymbol{h}_1^{\mathrm{H}} \boldsymbol{h}_2|^2 \left(\boldsymbol{h}_1^{\mathrm{H}} \left(\boldsymbol{I}_t + q_2 \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}}\right)^{-1} \boldsymbol{h}_1 + q_1 |\boldsymbol{h}_2^{\mathrm{H}} \left(\boldsymbol{I}_t + q_2 \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}}\right)^{-1} \boldsymbol{h}_1|^2\right) + \\
&+ \|\boldsymbol{h}_2\|_2^2 \left(1 + q_1 \boldsymbol{h}_1^{\mathrm{H}} \left(\boldsymbol{I}_t + q_2 \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}}\right)^{-1} \boldsymbol{h}_1\right)^2.
\end{aligned}$$

Application of the matrix inversion lemma on the factors $\left(\boldsymbol{I}_t + q_2 \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}}\right)^{-1}$ and multiplication on both sides of the equality with $(1 + q_2 \|\boldsymbol{h}_2\|_2^2)^2$ yields

$$\begin{aligned}
&\|\boldsymbol{h}_2\|_2^2 \left(1 + q_2 \|\boldsymbol{h}_2\|_2^2\right) \left(1 + q_1 \|\boldsymbol{h}_1\|_2^2\right) \left(\left(1 + q_2 \|\boldsymbol{h}_2\|_2^2\right) \left(1 + q_1 \|\boldsymbol{h}_1\|_2^2\right) + q_1 q_2 |\boldsymbol{h}_1^{\mathrm{H}} \boldsymbol{h}_2|^2\right) = \\
&\quad \|\boldsymbol{h}_2\|_2^2 \left(\left(1 + q_2 \|\boldsymbol{h}_2\|_2^2\right) \left(1 + q_1 \|\boldsymbol{h}_1\|_2^2\right) + q_1 q_2 |\boldsymbol{h}_1^{\mathrm{H}} \boldsymbol{h}_2|^2\right)^2 + \\
&\quad + q_1 q_2 |\boldsymbol{h}_1^{\mathrm{H}} \boldsymbol{h}_2|^2 \left(\left(\underbrace{\|\boldsymbol{h}_1\|_2^2 \left(1 + q_2 \|\boldsymbol{h}_2\|_2^2\right) - q_2 |\boldsymbol{h}_1^{\mathrm{H}} \boldsymbol{h}_2|^2}\right) \left(1 + q_2 \|\boldsymbol{h}_2\|_2^2\right) + q_1 |\boldsymbol{h}_1^{\mathrm{H}} \boldsymbol{h}_2|^2\right).
\end{aligned}$$

From here, the result follows immediately after replacing the factor marked by the brace in the second term on the right-hand side by the right-hand side of the optimality condition given by Eq. 2.31.

2.2.2.4 The capacity region

In this last section of the chapter we give a brief account of some of the most significant results on the information theoretical analysis of the Gaussian broadcast channel, which, eventually, led to the finding that the dirty paper coding region as defined by Eq. 2.24 is in fact the capacity region of the general Gaussian broadcast channel.

The idea of applying dirty paper coding to the broadcast channel with single receive antennas by successively encoding the users in the network is due to Caire et al. [23, 21, 22]. These authors also proved that for the two-user channel with single receive antennas, this transmission scheme is able to achieve the sum capacity. In order to prove that, they showed that the optimum sum rate achievable by using the DPC scheme is equal to the Sato upper bound on sum capacity given by Eq. 2.5. In [145] Yu et al. presented achievable rates for the general Gaussian broadcast channel based on DPC. Independent work by Viswanath et al. [132], Vishwanath et al. [128, 127] and Yu et al. [144, 143] showed that DPC achieves the sum capacity of the Gaussian broadcast channel with an arbitrary number of users and antennas. The first two authors based their proofs on duality between downlink and uplink and showed that maximization of the sum rate in the uplink is equivalent to the minimization problem that must be solved in order to compute the Sato bound. Yu et al. considered a point to point channel with a minimum mean squared error (MMSE) decision-feedback equalizer at the receiver. They first proved that, if statistically independent signals are transmitted, the feedback part of the filter can be moved to the transmitter by considering dirty paper coding without incurring any performance loss.

Then, they showed that for the worst case noise, i.e., the noise that achieves the Sato bound, the forward filter decomposes into a set of filters, one for each signal, and, therefore, cooperation becomes immaterial.

Independent work by Vishwanath et al. [129] and Tse et al. [122] showed that, conditioned on the use of Gaussian alphabets, the dirty paper coding region is optimum. In both works, this was proved by considering the degraded Gaussian vector channels that can be obtained by assuming a certain ordering of the users and providing a certain user with the outputs of all preceding users. Assuming Gaussian inputs, the capacity region of a degraded channel constructed in this way depends on the particular ordering of the users and on the correlation between noise processes of different users. However, for any choice of these parameters, it is obvious that the resulting region comprises the capacity region that can be achieved on the original channel with Gaussian alphabets. In order to tighten this outer bound the authors in [129], [122] considered the intersection of the capacity regions of all degraded channels that can be defined with arbitrary choices of noise correlations and user orderings. Based on BC-MAC duality, they were able to show that the resulting region is included in the dirty paper coding region of the original channel. However, optimality of Gaussian inputs could not be demonstrated in these works. Note that, as mentioned in Section 2.2.1, optimality of Gaussian inputs for the Gaussian scalar degraded broadcast channel was shown by Bergmans in [5], however, his proof does not easily extend to vector channels. In [134] Weingarten et al. gave a proof for the optimality of Gaussian inputs for degraded Gaussian vector channels, which combined with the results from [129], [122], finally proved that the DPC region is actually the capacity region of the Gaussian broadcast channel. The same authors gave a more direct proof of the optimality of the DPC region in [135] that does not rely on the tools employed in [129] and [122].

3 Optimization criteria and optimum approaches

In this chapter three design criteria are analyzed that represent different ways of selecting the operational point of a point to multipoint network out of the set of achievable rates. The derivation of the algorithms that achieve optimality for each of these criteria requires the understanding of basic optimization theoretic concepts such as Lagrangian duality or subgradients. A brief summary of the most important definitions and results is given in Appendix A.3.

3.1 Sum-rate maximization

The first design criterion that we shall discuss in this chapter is sum-rate maximization. The goal is to maximize the total amount of information per channel use that the transmitter sends to the receivers in the network. This criterion entirely neglects quality of service requirements that users might have. Instead, it will generally favor users with good channels against users with bad channels for the sake of overall throughput. However, there are certain scenarios where this criterion might be of practical interest. For instance, for symmetric broadcast channels, where the sum-rate maximization criterion generally leads to operational points at which all users get a similar amount of resources. But also in scenarios where the set of active users strongly fluctuates, this criterion might be a good choice for delay insensitive applications. In such scenarios, every user requesting transmission rate will be served in relatively short time with high probability due to the strong fluctuation in the set of competing users. Hotspots delivering internet services to a large number of users that intermittently request access to content might be one example of this kind of scenarios.

Assume a Gaussian broadcast channel as given by Eq. 2.6. As we have seen in the previous chapter, for the Gaussian broadcast channel, successive encoding based on dirty paper coding is optimum in the sense that all rates in the capacity region can be achieved with this scheme and time-sharing. Let π be a permutation function defining the order in which information for the different users in the network is encoded. Mathematically, we can state the sum-rate maximization problem as follows,

$$\max_{\pi, \boldsymbol{\Sigma}_{1,\ldots,K}} \sum_{k=1}^{K} R_k \left(\pi, \boldsymbol{\Sigma}_{1,\ldots,K}\right),$$

subject to $\sum_{k=1}^{K} \mathrm{Tr}\left\{\boldsymbol{\Sigma}_k\right\} \leq P$ and $\boldsymbol{\Sigma}_k \geq \mathbf{0}$, $\forall k$, where (cf. Eqn 2.19)

$$R_{\pi(k')} \left(\pi, \boldsymbol{\Sigma}_{1,\ldots,K}\right) = \log_2 \left(\frac{|\boldsymbol{I}_{r_{\pi(k')}} + \boldsymbol{H}_{\pi(k')} \sum_{k \geq k'} \boldsymbol{\Sigma}_{\pi(k)} \boldsymbol{H}_{\pi(k')}^{\mathrm{H}}|}{|\boldsymbol{I}_{r_{\pi(k')}} + \boldsymbol{H}_{\pi(k')} \sum_{k > k'} \boldsymbol{\Sigma}_{\pi(k)} \boldsymbol{H}_{\pi(k')}^{\mathrm{H}}|} \right). \tag{3.1}$$

For a fixed π, $R_k\left(\pi, \boldsymbol{\Sigma}_{1,\ldots,K}\right)$ is, in general, neither a concave nor a convex function of the matrices $\boldsymbol{\Sigma}_{1,\ldots,K}$ and the same holds for the sum of these rates. Thus, the sum-rate maximization problem stated in this way does not qualify as a convex optimization problem. The significance of convex optimization problems resides in the possibility of finding the global optimum by iteratively performing local searches. Fortunately, it turns out that, in this case, the apparent non-convexity is not an impediment for a successful computation of the global optimum based on local search algorithms. In fact, invoking the duality result between BC and MAC discussed in the last chapter, an equivalent convex optimization problem can be formulated as follows,

$$\max_{\bar{\pi}, \boldsymbol{Q}_{1,\ldots,K}} \sum_{k=1}^{K} R_k\left(\bar{\pi}, \boldsymbol{Q}_{1,\ldots,K}\right), \tag{3.2}$$

subject to $\sum_{k=1}^{K} \mathrm{Tr}\left\{\boldsymbol{Q}_k\right\} \leq P$ and $\boldsymbol{Q}_k \geq \boldsymbol{0}$, $\forall k$, where $\bar{\pi}$ is a permutation function indicating the order in which users are decoded and

$$R_{\bar{\pi}(k')}\left(\bar{\pi}, \boldsymbol{Q}_{1,\ldots,K}\right) = \log_2\left(\frac{|\boldsymbol{I}_t + \sum_{k \geq k'} \boldsymbol{H}_{\bar{\pi}(k)}^{\mathrm{H}} \boldsymbol{Q}_{\bar{\pi}(k)} \boldsymbol{H}_{\bar{\pi}(k)}|}{|\boldsymbol{I}_t + \sum_{k > k'} \boldsymbol{H}_{\bar{\pi}(k)} \boldsymbol{Q}_{\bar{\pi}(k)} \boldsymbol{H}_{\bar{\pi}(k)}^{\mathrm{H}}|}\right). \tag{3.3}$$

By substituting Eq. 3.3 in Eq. 3.2, the objective function of this optimization problem can explicitly be written as

$$\sum_{k=1}^{K} R_k\left(\bar{\pi}, \boldsymbol{Q}_{1,\ldots,K}\right) = \log_2\left(\left|\boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{Q}_k \boldsymbol{H}_k\right|\right). \tag{3.4}$$

We observe that even though the rate of each particular user does depend on the decoding order $\bar{\pi}$, the sum of rates is independent of $\bar{\pi}$. Furthermore, the sum of rates is a strictly convex function of the transmit covariance matrices $\boldsymbol{Q}_{1,\ldots,K}$ even if the individual rates are not. Thus, in the dual MAC the sum-rate maximization problem under an overall transmit power constraint is a convex optimization problem. In addition, the optimum can be achieved with any decoding order. Of course, in the broadcast channel the optimum is achieved by any set of transmit covariance matrices computed from the optimum $\boldsymbol{Q}_{1,\ldots,K}$ by fixing a decoding order and using the duality transformations given in [128]. The optimum encoding order is obtained by reversing the decoding order fixed in the dual MAC for computation of the BC transmit covariance matrices. Due to the strict concavity of $\log|\bullet|$ there is a unique set of transmit covariance matrices that maximize the sum rate in the MAC. However, due to the polymatroid structure of the MAC region corresponding to fixed statistics, optimality is shared by the set of rate vectors comprised by the convex polytope whose vertices are determined by the $K!$ possible decoding orders. Following the discussion in Section 2.2.2.3, in the BC, each of these vertices is achieved by a different set of transmit covariance matrices and DPC-based successive encoding. Conjecture 2.2.3 suggests that, in the BC, these vertices might also be achievable with just one set of covariance matrices by employing non-DPC-based successive encoding schemes.

3.1.1 Memoryless channels

In this section we present an overview of the algorithmic approaches that have been proposed in the literature in order to solve Problem 3.2 efficiently. In the next section we will specialize some of these approaches to channels with block diagonal structure, which typically arise when using OFDM as a transmission scheme in order to combat multipath in time-dispersive channels.

Before starting the discussion on algorithmic schemes that solve Problem 3.2, let us consider the sum-rate optimization problem in the MAC with individual power constraints, i.e.,

$$\max_{\boldsymbol{Q}_{1,\ldots,K}} \log_2 \left(\left| \boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{Q}_k \boldsymbol{H}_k \right| \right), \tag{3.5}$$

subject to $\boldsymbol{Q}_k \geq \boldsymbol{0}$ and $\mathrm{Tr}\{\boldsymbol{Q}_k\} \leq P_k, \forall k$. For this optimization problem, Yu et al. presented an algorithmic solution in [146, 147] that was named iterative waterfilling. This algorithm is the base upon which two of the most significant algorithmic solutions to Problem 3.2 are built. The iterative waterfilling algorithm, which solves Problem 3.5 iteratively, is based on the following observation. Assume that after the ℓth iteration we have matrices $\boldsymbol{Q}_{1,\ldots,K}^{(\ell)}$. Increase of the objective function in the $\ell+1$ iteration is guaranteed by selecting a user, say j, letting the covariance matrices of all other users unchanged, i.e., $\boldsymbol{Q}_k^{(\ell+1)} = \boldsymbol{Q}_k^{(\ell)}$, $k \neq j$, and computing the new covariance matrix for user j as

$$\boldsymbol{Q}_j^{(\ell+1)} = \arg\max_{\boldsymbol{Q}_j} \log_2 \left(\left| \boldsymbol{I}_t + \sum_{k \neq j} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{Q}_k^{(\ell)} \boldsymbol{H}_k + \boldsymbol{H}_j^{\mathrm{H}} \boldsymbol{Q}_j \boldsymbol{H}_j \right| \right),$$

subject to $\boldsymbol{Q}_j \geq \boldsymbol{0}$ and $\mathrm{Tr}\left\{\boldsymbol{Q}_j\right\} \leq P_j$. That this step necessarily leads to an improvement becomes apparent if the objective function is rewritten as

$$\begin{aligned} &\log_2 \left(\left| \boldsymbol{I}_t + \sum_{k \neq j} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{Q}_k^{(\ell)} \boldsymbol{H}_k + \boldsymbol{H}_j^{\mathrm{H}} \boldsymbol{Q}_j \boldsymbol{H}_j \right| \right) = \\ &= \log_2 \left(\left| \boldsymbol{I}_t + \sum_{k \neq j} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{Q}_k^{(\ell)} \boldsymbol{H}_k \right| \right) + \log_2 \left(\left| \boldsymbol{I}_t + \left(\boldsymbol{I}_t + \sum_{k \neq j} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{Q}_k^{(\ell)} \boldsymbol{H}_k \right)^{-1} \boldsymbol{H}_j^{\mathrm{H}} \boldsymbol{Q}_j \boldsymbol{H}_j \right| \right). \end{aligned}$$

Being user j the first encoded user, the first term on the right-hand side of this expression represents the sum rate achieved by all other users. This term is independent of the covariance matrix of user j. The second term is the rate achieved by user j. Certainly, choosing $\boldsymbol{Q}_j$ such that the second term is maximized leads to an improvement of the total sum-rate. The KKT optimality conditions corresponding to this maximization yield

$$\hat{\boldsymbol{H}}_j \left(\boldsymbol{I}_t + \hat{\boldsymbol{H}}_j^{\mathrm{H}} \boldsymbol{Q}_j \hat{\boldsymbol{H}}_j \right)^{-1} \hat{\boldsymbol{H}}_j^{\mathrm{H}} + \boldsymbol{\Phi}_j - \mu_j \boldsymbol{I}_{r_j} = 0, \tag{3.6}$$

where

$$\hat{\boldsymbol{H}}_j = \boldsymbol{H}_j \left(\boldsymbol{I}_t + \sum_{k \neq j} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{Q}_k^{(\ell)} \boldsymbol{H}_k \right)^{-1/2}$$

is the effective channel seen by user j, $\boldsymbol{\Phi}_j \geq \mathbf{0}$ is the Lagrange multiplier associated with the constraint $\boldsymbol{Q}_j \geq \mathbf{0}$ and $\mu_j \geq 0$ is the Lagrange multiplier corresponding to the transmit power constraint $\mathrm{Tr}\left\{\boldsymbol{Q}_j\right\} \leq P_j$. It can be easily shown that the matrix $\boldsymbol{Q}_j$ that fulfills this condition has the same eigenvectors as $\hat{\boldsymbol{H}}_j \hat{\boldsymbol{H}}_j^{\mathrm{H}}$ and eigenvalues according to a waterfilling power distribution on the eigenvalues of this matrix [119]. In each iteration, the matrix of a different user is updated by keeping the matrices of all other users fixed. In this way, the value of the objective function is increased until a fixed point is achieved in which every user waterfills its power over the effective channel determined by its own channel and the interference from all other users. The pseudocode for this approach is given in Algorithm 3.1.

Algorithm 3.1 Iterative waterfilling

$\boldsymbol{Q}_k^{(0)} \leftarrow \mathbf{0}, \quad k = 1, \ldots, K, \quad \ell = 0$
$R^{\mathrm{new}} \leftarrow 0$
repeat
  $R^{\mathrm{old}} \leftarrow R^{\mathrm{new}}$
  for $j = 1$ to K **do**
    $\boldsymbol{Z} \leftarrow \boldsymbol{I}_t + \sum_{k \neq j} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{Q}_k^{(\ell)} \boldsymbol{H}_k$
    $\boldsymbol{Q}_j^{(\ell+1)} \leftarrow \underset{\boldsymbol{Q}_j}{\arg\max} \ \log_2\left(\left|\boldsymbol{Z} + \boldsymbol{H}_j^{\mathrm{H}} \boldsymbol{Q}_j \boldsymbol{H}_j\right|\right)$
    subject to $\boldsymbol{Q}_j \geq \mathbf{0}$ and $\mathrm{Tr}\left\{\boldsymbol{Q}_j\right\} \leq P_j$
    $\boldsymbol{Q}_k^{(\ell+1)} \leftarrow \boldsymbol{Q}_k^{(\ell)}, \ k \neq j$
    $\ell \leftarrow \ell + 1$
  end for
  $R^{\mathrm{new}} \leftarrow \log_2\left(\left|\boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{Q}_k^{(\ell)} \boldsymbol{H}_k\right|\right)$
until $R^{\mathrm{new}} - R^{\mathrm{old}} < \epsilon$

3.1.1.1 Sum-power iterative waterfilling

If the individual power constraints in Problem 3.5 are replaced by a total power constraint, convergence of Algorithm 3.1 can no longer be guaranteed. In this case, users are coupled by the power constraint and, hence, optimizing the transmit strategy of a particular user at a time does not necessarily lead to an increase of the objective function. The algorithmic approach taken in [65, 130] in order to find a solution consists of restating Problem 3.2 in an equivalent form but with decoupled constraints. This allows a straightforward application of the basic principles behind Algorithm 3.1.

Consider the following optimization problem

$$\max_{\substack{\boldsymbol{Q}_{1,\ldots,K} \\ 1,\ldots,K}} \sum_{j=1}^{K} \frac{1}{K} \log_2\left(\left|\boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{Q}_{j,k} \boldsymbol{H}_k\right|\right), \tag{3.7}$$

subject to $\boldsymbol{Q}_{j,k} \geq \mathbf{0}$, $\forall j, k$, and $\sum_{k=1}^{K} \mathrm{Tr}\left\{\boldsymbol{Q}_{[j+k]_K,k}\right\} \leq P$, $j = 1, \ldots, K$, where $[x]_K = \mathrm{mod}\left((x-1), K\right) + 1$. It is obvious that any value achievable by the objective function

of Problem 3.2 can also be reached by the objective function of this problem by choosing $\boldsymbol{Q}_{1,k} = \cdots = \boldsymbol{Q}_{K,k} = \boldsymbol{Q}_k$, $\forall k$. On the other hand, due to concavity of $\log|\bullet|$,

$$\sum_{j=1}^{K} \frac{1}{K} \log_2 \left(\left| \boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{Q}_{j,k} \boldsymbol{H}_k \right| \right) \leq \log_2 \left(\left| \boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_k^{\mathrm{H}} \hat{\boldsymbol{Q}}_k \boldsymbol{H}_k \right| \right), \tag{3.8}$$

where $\hat{\boldsymbol{Q}}_k = \sum_{j=1}^{K} \boldsymbol{Q}_{j,k}/K$. Thus, any value achievable by the objective function of Problem 3.7 is also reachable by the objective function of 3.2, which shows the equivalence of these two problems. Note that in Problem 3.7, for any $1 \leq i \leq K$, the transmit covariance matrices $\boldsymbol{Q}_{[i+k]_K,k}$, $k = 1, ..., K$, are decoupled of matrices $\boldsymbol{Q}_{[j+k]_K,k}$, $k = 1, ..., K$, $j \neq i$, as they are subject to different power constraints. Therefore, an iterative algorithm can be applied that increases the objective function in each iteration by optimizing over a set of covariance matrices $\boldsymbol{Q}_{[i+k]_K,k}$, $k = 1, ..., K$, while keeping all other covariance matrices $\boldsymbol{Q}_{[j+k]_K,k}$, $k = 1, ..., K$, $j \neq i$, fixed. Iterations are repeated until a fixed point is reached (see Algorithm 3.2). Due to Eq. 3.8, $\boldsymbol{Q}_{1,k} = \cdots = \boldsymbol{Q}_{K,k} = \boldsymbol{Q}_k$, $\forall k$, must hold at this point.

Algorithm 3.2 Sum-power iterative waterfilling (cyclic coordinate ascent)

1: $\boldsymbol{Q}_{j,k} \leftarrow \boldsymbol{0}, \quad k = 1, \ldots, K, \ j = 1, \ldots, K, \quad i = 1$
2: $R^{\mathrm{new}} \leftarrow 0$
3: **repeat**
4: $\quad R^{\mathrm{old}} \leftarrow R^{\mathrm{new}}$
5: $\quad$ **for** $m = 1$ to K **do**
6: $\quad\quad \boldsymbol{Z}_{[i+m]_K} \leftarrow \boldsymbol{I}_t + \sum_{k \neq m} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{Q}_{[i+m]_K,k} \boldsymbol{H}_k$
7: $\quad$ **end for**
8: $\quad \left\{ \boldsymbol{Q}_{[i+k]_K,k} | k = 1, \ldots K \right\} \leftarrow \underset{\boldsymbol{Q}_{1,\ldots,K}}{\arg\max} \ \frac{1}{K} \sum_{k=1}^{K} \log_2 \left(\left| \boldsymbol{Z}_{[i+k]_K} + \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{Q}_k \boldsymbol{H}_k \right| \right)$
$\quad$ subject to $\boldsymbol{Q}_k \geq \boldsymbol{0}$, $\forall k$, and $\sum_{k=1}^{K} \mathrm{Tr}\{\boldsymbol{Q}_k\} \leq P$
9: $\quad i \leftarrow [i+1]_K$
10: $\quad R^{\mathrm{new}} \leftarrow \frac{1}{K} \sum_{j=1}^{K} \log_2 \left(\left| \boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{Q}_{j,k} \boldsymbol{H}_k \right| \right)$
11: **until** $R^{\mathrm{new}} - R^{\mathrm{old}} < \epsilon$

A drawback of Algorithm 3.2 is that $K(K-1)$ covariance matrices need to be stored in memory for the execution of one iteration. A reduction of storage can be achieved by including the following averaging step after line 8,

$$\boldsymbol{Q}_{m,k} \leftarrow \frac{1}{K} \sum_{j=1}^{K} \boldsymbol{Q}_{j,k}, \quad k = 1, \ldots, K, \ m = 1, \ldots, K.$$

Due to Eq. 3.8, this step leads to an improvement of the objective function. Furthermore, the transmit covariance matrices are no longer a function of index m and, therefore, only K matrices need to be stored. However, a negative side effect of the averaging step is a reduced convergence rate. A compact pseudocode for this modified algorithm is given in Algorithm 3.3.

Algorithm 3.3 Sum-power iterative waterfilling

1: $\boldsymbol{Q}_k^{(0)} \leftarrow \boldsymbol{0}, \quad k = 1, \ldots, K, \quad \ell = 0$
2: $R^{\text{new}} \leftarrow 0$
3: **repeat**
4: $\quad R^{\text{old}} \leftarrow R^{\text{new}}$
5: $\quad$ **for** $m = 1$ to K **do**
6: $\quad\quad \boldsymbol{Z}_m \leftarrow \boldsymbol{I}_t + \sum_{k \neq m} \boldsymbol{H}_k^{\text{H}} \boldsymbol{Q}_k^{(\ell)} \boldsymbol{H}_k$
7: $\quad$ **end for**
8: $\quad \boldsymbol{M}_{1,\ldots,K} \leftarrow \underset{\boldsymbol{Q}_{1,\ldots,K}}{\arg\max} \ \frac{1}{K} \sum_{k=1}^{K} \log_2 \left(\left| \boldsymbol{Z}_k + \boldsymbol{H}_k^{\text{H}} \boldsymbol{Q}_k \boldsymbol{H}_k \right| \right)$
$\quad$ subject to $\boldsymbol{Q}_k \geq \boldsymbol{0}$, $\forall k$, and $\sum_{k=1}^{K} \text{Tr}\{\boldsymbol{Q}_k\} \leq P$
9: $\quad \boldsymbol{Q}_k^{(\ell+1)} \leftarrow \frac{1}{K} \left(\boldsymbol{M}_k + (K-1) \boldsymbol{Q}_k^{(\ell)} \right)$
10: $\quad \ell \leftarrow \ell + 1$
11: $\quad R^{\text{new}} \leftarrow \log_2 \left(\left| \boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_k^{\text{H}} \boldsymbol{Q}_k^{(\ell)} \boldsymbol{H}_k \right| \right)$
12: **until** $R^{\text{new}} - R^{\text{old}} < \epsilon$

3.1.1.2 Dual decomposition

A second approach to the computation of sum-rate maximizing transmit covariance matrices is due to Yu and was presented in [141, 142]. In this approach, auxiliary power variables are introduced for each user in order to restate Problem 3.2 as follows,

$$\max_{\boldsymbol{Q}_{1,\ldots,K}, P_{1,\ldots,K}} \log_2 \left(\left| \boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_k^{\text{H}} \boldsymbol{Q}_k \boldsymbol{H}_k \right| \right), \tag{3.9}$$

subject to $\boldsymbol{Q}_k \geq \boldsymbol{0}$, $\text{Tr}\{\boldsymbol{Q}_k\} \leq P_k$, $\forall k$, and $\sum_{k=1}^{K} P_k \leq P$. The Lagrangian function of this optimization problem with respect to the last constraint can be written as

$$L\left(\boldsymbol{Q}_{1,\ldots,K}, P_{1,\ldots,K}, \lambda\right) = \log_2 \left(\left| \boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_k^{\text{H}} \boldsymbol{Q}_k \boldsymbol{H}_k \right| \right) + \lambda \left(P - \sum_{k=1}^{K} P_k \right).$$

In turn, the corresponding Lagrangian dual function is defined as

$$g(\lambda) = \max_{\boldsymbol{Q}_{1,\ldots,K}, P_{1,\ldots,K}} L\left(\boldsymbol{Q}_{1,\ldots,K}, P_{1,\ldots,K}, \lambda\right), \tag{3.10}$$

subject to $\boldsymbol{Q}_k \geq \boldsymbol{0}$, $\text{Tr}\{\boldsymbol{Q}_k\} \leq P_k$, $\forall k$. The dual problem itself reads

$$\min_{\lambda} \ g(\lambda), \tag{3.11}$$

subject to $\lambda \geq 0$. Problem 3.9 is convex. Furthermore, the set of feasible transmit covariance matrices and individual powers has a non-empty interior. Therefore, Slater's constraint qualification is satisfied and strong duality holds [13], i.e., the maximum of Problem 3.9 coincides with the minimum of Problem 3.11. This property allows for an algorithmic

approach that iteratively searches for the solution to Problem 3.11 instead of directly solving Problem 3.9. Due to the fact that $g(\lambda)$ is generally non-differentiable the search is done relying on a subgradient method.

Assume $\lambda^{(\ell)}$ to be the value obtained for λ after the ℓth iteration. Computation of $g(\lambda^{(\ell)})$ is done by solving Problem 3.10. The KKT conditions of this problem yield

$$\hat{\boldsymbol{H}}_j \left(\boldsymbol{I}_t + \hat{\boldsymbol{H}}_j^{\mathrm{H}} \boldsymbol{Q}_j \hat{\boldsymbol{H}}_j \right)^{-1} \hat{\boldsymbol{H}}_j^{\mathrm{H}} + \boldsymbol{\Phi}_j - \lambda^{(\ell)} \boldsymbol{I}_{r_j} = \boldsymbol{0}, \quad \forall j \tag{3.12}$$

where

$$\hat{\boldsymbol{H}}_j = \boldsymbol{H}_j \left(\boldsymbol{I}_t + \sum_{k \neq j} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{Q}_k \boldsymbol{H}_k \right)^{-1/2}$$

is the effective channel seen by user j and $\boldsymbol{\Phi}_j \geq \boldsymbol{0}$ is the Lagrange multiplier associated with the constraint $\boldsymbol{Q}_j \geq \boldsymbol{0}$. Note that the conditions given by Eqs. 3.12 are identical to the condition given by Eq. 3.6. Thus, as in Problem 3.5, here optimality is also achieved when all users transmit their signals aligned with the eigenvectors of their respective effective channels with a waterfilling distribution of the powers. While in Problem 3.5 the water level, determined by μ_k, is different for every user, in this problem the water level, determined by $\lambda^{(\ell)}$, is the same for every user. As in Problem 3.5, this fixed point can be achieved by updating the covariance matrix of one user at a time while keeping the covariance matrices of all other users fixed. That is, Algorithm 3.1 can be readily used in order to find the solution to Problem 3.10. The only difference is that in line 7 maximization must be performed subject to the water level dictated by $\lambda^{(\ell)}$ rather than to an individual power constraint.

Once $g(\lambda^{(\ell)})$ has been evaluated, a subgradient of this function at $\lambda^{(\ell)}$ must be found, which gives an appropriate direction for computation of $\lambda^{(\ell+1)}$. Recalling the definition of subgradient in Eq. A.11, it can be easily shown that a subgradient of $g(\lambda)$ at $\lambda^{(\ell)}$ is given by

$$s^{(\ell)} = \left(P - \sum_{k=1}^{K} P_k^{(\ell)} \right),$$

where $P_{1,\ldots,K}^{(\ell)}$ are maximizers of Problem 3.10 for $\lambda = \lambda^{(\ell)}$. Indeed,

$$\begin{aligned} g(\lambda^{(\ell)} + \Delta\lambda) \; - \; & g(\lambda^{(\ell)}) = \\ = \; & \max_{\boldsymbol{Q}_{1,\ldots,K}, P_{1,\ldots,K}} L\left(\boldsymbol{Q}_{1,\ldots,K}, P_{1,\ldots,K}, \lambda^{(\ell)} + \Delta\lambda\right) - L\left(\boldsymbol{Q}_{1,\ldots,K}^{(\ell)}, P_{1,\ldots,K}^{(\ell)}, \lambda^{(\ell)}\right) \\ \geq \; & L\left(\boldsymbol{Q}_{1,\ldots,K}^{(\ell)}, P_{1,\ldots,K}^{(\ell)}, \lambda^{(\ell)} + \Delta\lambda\right) - L\left(\boldsymbol{Q}_{1,\ldots,K}^{(\ell)}, P_{1,\ldots,K}^{(\ell)}, \lambda^{(\ell)}\right) \\ = \; & \Delta\lambda \left(P - \sum_{k=1}^{K} P_k^{(\ell)} \right), \end{aligned}$$

where $\boldsymbol{Q}_{1,\ldots,K}^{(\ell)}$, $P_{1,\ldots,K}^{(\ell)}$ are maximizers of Problem 3.10 for $\lambda = \lambda^{(\ell)}$. Thus, a convenient search direction is given by $-s^{(\ell)}$, i.e., if the constraint is violated, $\lambda^{(\ell)}$ should be increased. Otherwise, $\lambda^{(\ell)}$ should be decreased. Due to the monotonicity of the constraint with respect

to λ, a bisection method can be used in order to choose the step size of each update until the constraint is fulfilled with equality within a certain precision. Once the optimum λ is found, the maximizers of Problem 3.10 turn out to be the maximizers of Problem 3.9. This can be easily shown by observing,

$$\begin{aligned}\log_2\left(\left|\boldsymbol{I}_t+\sum_{k=1}^{K}\boldsymbol{H}_k^{\mathrm{H}}\bar{\boldsymbol{Q}}_k\boldsymbol{H}_k\right|\right) &= g(\bar{\lambda})\\ &\geq L\left(\bar{\boldsymbol{Q}}_{1,\ldots,K},\bar{P}_{1,\ldots,K},\bar{\lambda}\right)\\ &\geq \log_2\left(\left|\boldsymbol{I}_t+\sum_{k=1}^{K}\boldsymbol{H}_k^{\mathrm{H}}\bar{\boldsymbol{Q}}_k\boldsymbol{H}_k\right|\right),\end{aligned}$$

where $\bar{\boldsymbol{Q}}_{1,\ldots,K}$, $\bar{P}_{1,\ldots,K}$ and $\bar{\lambda}$ are maximizers of the primal and the dual problems, respectively. The first equation is due to strong duality. The first inequality is due to the definition of $g(\bar{\lambda})$ as the maximum of $L\left(\boldsymbol{Q}_{1,\ldots,K},P_{1,\ldots,K},\bar{\lambda}\right)$. Finally, the second inequality is due to the fact that the Lagrangian function with non-negative multipliers is always greater than the corresponding objective function with feasible arguments. The pseudocode for this procedure is given in Algorithm 3.4. The term dual decomposition alludes to the fact that the link between the transmit covariance matrices, established by the total power constraint in Problem 3.2, is effectively "decomposed" if we consider the dual problem in order to find the solution.

Algorithm 3.4 Sum-rate maximization via dual decomposition

1: Initialize $\lambda_{\min}$ and $\lambda_{\max}$
2: $\lambda^{(0)}=(\lambda_{\max}+\lambda_{\min})/2$, $\ell\leftarrow 0$
3: **repeat**
4: Compute $g(\lambda^{(\ell)})$ using Algorithm 3.1
5: **if** $\left(P-\sum_{k=1}^{K}P_k^{(\ell)}\right)>0$ **then**
6: $\lambda^{(\ell+1)}\leftarrow\frac{\lambda^{(\ell)}+\lambda_{\min}}{2}$
7: $\lambda_{\max}\leftarrow\lambda^{(\ell)}$
8: **else**
9: $\lambda^{(\ell+1)}\leftarrow\frac{\lambda^{(\ell)}+\lambda_{\max}}{2}$
10: $\lambda_{\min}\leftarrow\lambda^{(\ell)}$
11: **end if**
12: $\ell\leftarrow\ell+1$
13: **until** $\lambda_{\max}-\lambda_{\min}<\epsilon$

3.1.1.3 Further work

A number of algorithms have been proposed in recent years in order to speed up convergence or reduce complexity of both the sum-power iterative waterfilling and the dual decomposition algorithm reviewed in previous sections. In [11] the authors present a procedure to speed up the convergence of Algorithm 3.3 by optimizing the weights of the

averaging step in line 9. In [33] the authors report reduction in complexity and improvement in convergence speed with respect to Algorithm 3.3 by just selecting two users at random in order to update their covariance matrices while keeping the matrices of all other users fixed in each iteration. The issue of how precise $g(\lambda)$ must be computed in line 4 of Algorithm 3.4 in order to reduce the number of iterations but still guarantee convergence is addressed in [34]. Adopting a completely different approach, in [63], the authors proposed an algorithm that finds optimum precoding matrices rather than transmit covariance matrices. This algorithm is based on a projected gradient search and has the advantageous property that no eigenvalue decompositions are needed.

3.1.2 Time-dispersive channels

As discussed in Section 1.3, OFDM can be employed for transmission over time-dispersive channels in order to effectively transform the multipath channel into a set of decoupled, non-dispersive channels. In a point to multipoint transmission setting, the model given by Eq. 2.6 applies to each of the channels of the MIMO OFDM model given by Eq. 1.5. In particular, the signal received by user $k \in \{1, \ldots, K\}$ on subcarrier $n \in \{1, \ldots, N\}$ can be written as

$$\boldsymbol{y}_{k,n} = \boldsymbol{H}_{k,n}\boldsymbol{x}_n + \boldsymbol{n}_{k,n}.$$

This system model can be reduced to a memoryless multiuser MIMO model by writing

$$\boldsymbol{y}_k = \boldsymbol{H}_k\boldsymbol{x} + \boldsymbol{n}_k, \quad k = 1, \ldots, K, \tag{3.13}$$

where $\boldsymbol{y}_k = [\ \boldsymbol{y}_{k,1}^{\mathrm{T}} \ \cdots \ \boldsymbol{y}_{k,N}^{\mathrm{T}}\ ]^{\mathrm{T}}$, $\boldsymbol{n}_k = [\ \boldsymbol{n}_{k,1}^{\mathrm{T}} \ \cdots \ \boldsymbol{n}_{k,N}^{\mathrm{T}}\ ]^{\mathrm{T}}$, $\boldsymbol{x} = [\ \boldsymbol{x}_1^{\mathrm{T}} \ \cdots \ \boldsymbol{x}_N^{\mathrm{T}}\ ]^{\mathrm{T}}$ and

$$\boldsymbol{H}_k = \begin{bmatrix} \boldsymbol{H}_{k,1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{H}_{k,2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{H}_{k,N} \end{bmatrix}. \tag{3.14}$$

The essential difference between this system model and the model for general memoryless MIMO broadcast channels (cf. Eqn 2.6) resides in the fact that the channel matrices $\boldsymbol{H}_k \in \mathbb{C}^{Nr_k \times Nt}$ are block diagonal in this case. Nevertheless, from a theoretical point of view, the multiuser MIMO OFDM channel can be viewed as a high dimensional memoryless broadcast channel for which all results discussed in Section 2.2 apply. Also, the sum-rate maximization problem statement in Eq. 3.2 directly applies to this setting. An obvious approach to the computation of the sum-rate optimum transmit strategy for this model is the direct application of the algorithms discussed in the previous section. Although theoretically sound, this approach shows a crucial disadvantage. The search space is given by the Cartesian product of K Hermitian matrices of dimensions $Nr_k \times Nr_k$, $k = 1, \ldots, K$. Taking into account that all the known algorithms have a cubic order of complexity in the number of rows or columns of the covariance matrices per iteration [63], we observe that the direct application of the sum-rate maximizing algorithms for memoryless channels to the MIMO OFDM setting yields a cubic complexity order in the number of subcarriers. Due to this fact, the computational power required to compute optimum

covariance matrices in this setting becomes already prohibitive for numbers of subcarriers far below those commonly used in practice. Fortunately, in the following we show that the optimum covariance matrices inherit the block diagonal structure of the channel matrices. As we shall see, this allows to design extensions of the algorithms discussed in the previous section that have linear complexity in the number of subcarriers.

To begin with, we rewrite the sum-rate maximization problem statement by replacing Eq. 3.4 in Eq. 3.2, i.e.,

$$\max_{\boldsymbol{Q}_{1,\ldots,K}} \log_2 \left(\left| \boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{Q}_k \boldsymbol{H}_k \right| \right), \tag{3.15}$$

subject to $\boldsymbol{Q}_k \geq \boldsymbol{0}$, $\forall k$, and $\sum_{k=1}^{K} \mathrm{Tr}\left\{\boldsymbol{Q}_k\right\} \leq P$. Let $\bar{\boldsymbol{Q}}_{1,\ldots,K}$ be the set of optimum covariance matrices for this problem and let $\bar{\boldsymbol{Q}}_k^{\mathrm{b}}$ be a block diagonal matrix obtained out of $\bar{\boldsymbol{Q}}_k$ by setting the off-diagonal elements to zero, i.e.,

$$\left[\bar{\boldsymbol{Q}}_k^{\mathrm{b}}\right]_{i,j} = \begin{cases} \left[\bar{\boldsymbol{Q}}_k\right]_{i,j}, & \left\lfloor \frac{i}{r_k} \right\rfloor = \left\lfloor \frac{j}{r_k} \right\rfloor \\ 0, & \left\lfloor \frac{i}{r_k} \right\rfloor \neq \left\lfloor \frac{j}{r_k} \right\rfloor \end{cases},$$

where $\lfloor \bullet \rfloor$ returns the largest integer strictly below the argument. Due to the block diagonal structure of the channel we can write

$$\boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_k^{\mathrm{H}} \bar{\boldsymbol{Q}}_k^{\mathrm{b}} \boldsymbol{H}_k = \left(\boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_k^{\mathrm{H}} \bar{\boldsymbol{Q}}_k \boldsymbol{H}_k \right)^{\mathrm{b}}, \tag{3.16}$$

where the matrix on the right-hand side is obtained by keeping the blocks of dimension $t \times t$ in the main diagonal of $\boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_k^{\mathrm{H}} \bar{\boldsymbol{Q}}_k \boldsymbol{H}_k$ and setting the off-diagonal elements to zero. Note also that the matrices $\bar{\boldsymbol{Q}}_{1,\ldots,K}^{\mathrm{b}}$ satisfy both the power constraint and the positive semidefinite constraint of Problem 3.15. In the following, we show that given a matrix $\boldsymbol{A} \in \mathbb{H}^{M \times M}$, $\boldsymbol{A} \geq \boldsymbol{0}$,

$$\log_2 \left(|\boldsymbol{A}| \right) \leq \log_2 \left(\left| \boldsymbol{A}^{\mathrm{b}} \right| \right), \tag{3.17}$$

where $\boldsymbol{A}^{\mathrm{b}}$ is a block diagonal matrix obtained by keeping blocks of any dimensions on main diagonal of $\boldsymbol{A}$ and setting the off-diagonal entries to zero. To this end, let $\mathbf{a}_{1,\ldots,S}$ be a set of random vectors with $\mathbf{a}_s \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{A}_s)$, $s = 1, \ldots, S$. In addition, let $\mathbf{a} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{A})$ be the random vector defined as $\mathbf{a} = \left[\begin{array}{ccc} \mathbf{a}_1^{\mathrm{T}} & \cdots & \mathbf{a}_S^{\mathrm{T}} \end{array} \right]^{\mathrm{T}}$. It holds

$$\begin{aligned} S \log \pi e + \log |\boldsymbol{A}| &= h(\mathbf{a}_1, \ldots, \mathbf{a}_S) = \sum_{s=1}^{S} h(\mathbf{a}_s | \mathbf{a}_1, \ldots, \mathbf{a}_{s-1}) \\ &\leq \sum_{s=1}^{N} h(\mathbf{a}_s) = S \log \pi e + \log \prod_{s=1}^{S} |\boldsymbol{A}_s| = S \log \pi e + \log |\boldsymbol{A}^{\mathrm{b}}|, \end{aligned}$$

where the inequality follows from the fact that conditioning reduces entropy and

$$\boldsymbol{A}^{\mathrm{b}} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{A}_S \end{bmatrix}.$$

This shows the validity of Eq. 3.17, which is a trivial generalization of the Hadamard inequality to block diagonal matrices [40]. Combining Eq. 3.17 and Eq. 3.16 we can write

$$\log_2\left(\left|\boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_k^{\mathrm{H}} \bar{\boldsymbol{Q}}_k \boldsymbol{H}_k\right|\right) \leq \log_2\left(\left|\boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_k^{\mathrm{H}} \bar{\boldsymbol{Q}}_k^{\mathrm{b}} \boldsymbol{H}_k\right|\right)$$

from which it follows that optimum covariance matrices are block diagonal. Hence, without loss of optimality, Problem 3.15 can be stated as

$$\max_{\substack{\boldsymbol{Q}_{1,\ldots,K} \\ 1,\ldots,N}} \sum_{n=1}^{N} \log_2\left|\boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_{k,n}^{\mathrm{H}} \boldsymbol{Q}_{k,n} \boldsymbol{H}_{k,n}\right|, \tag{3.18}$$

subject to $\sum_{k=1,n=1}^{K,N} \mathrm{Tr}\{\boldsymbol{Q}_{k,n}\} \leq P$ and $\boldsymbol{Q}_{k,n} \geq \boldsymbol{0}$, $\forall k, n$. Here, $\boldsymbol{Q}_{k,n} \in \mathbb{H}^{r_k \times r_k}$ is the transmit covariance matrix corresponding to user k on subcarrier n. The matrices $\boldsymbol{Q}_{k,n}$, $n = 1, \ldots, N$, are nothing else than the blocks on the diagonal of matrix $\boldsymbol{Q}_k$ in Problem 3.15. The off-diagonal blocks of this matrix disappear from the problem formulation as we know that they can be set to zero without loss of optimality. That is, the search space reduces to the Cartesian product of KN Hermitian matrices of dimension $r_k \times r_k$. Using the duality transformations given in [128] it can be shown that block diagonal matrices $\boldsymbol{Q}_{1,\ldots,K}$ in the MAC correspond to block diagonal matrices $\boldsymbol{\Sigma}_{1,\ldots,K}$ in the BC. Furthermore, the transmit covariance matrices $\boldsymbol{\Sigma}_{k,n}$, $k = 1, \ldots, K$, corresponding to subcarrier n in the BC are exclusively determined by the transmit covariance matrices $\boldsymbol{Q}_{k,n}$, $k = 1, \ldots, K$, corresponding to that subcarrier in the MAC, i.e., these matrices are independent of the MAC covariance matrices corresponding to any other subcarrier. Recalling the discussion preceding Section 3.1.1 for general MAC channels, sum-capacity can be achieved by any of the $K!$ possible decoding orders, which yields the $K!$ vertices of the polymatroid corresponding to the sum-rate optimum covariance matrices. In the general MAC, all other points within the polytope defined by these vertices can be achieved either by time-sharing or joint decoding (cf. Section 2.2.2.3). A special feature of the MIMO OFDM multiple access channel is that, apart from the vertices, there are other points within this polytope that are also achievable by performing successive decoding. These are all the rate vectors that are achieved by varying the decoding order across subcarriers. This is possible due to the fact that sum-capacity for this channel is achieved by transmitting statistically independent signals across subcarriers. Therefore, separate detection on each subcarrier is optimum. That is, on each subcarrier the detector can choose a decoding order without considering the detection order on any other subcarrier. Thus, the total number of sum-rate optimum rate vectors that can be achieved by performing successive decoding on the MIMO OFDM MAC amounts to $(K!)^N$. Correspondingly, assuming DPC-based successive encoding, there are $(K!)^N$ different transmission statistics that achieve sum-capacity in the dual MIMO OFDM BC.

3.1.2.1 Sum-power iterative waterfilling

The algorithmic approach followed in Section 3.1.1.1 to arrive at a solution for Problem 3.2 will be also taken here in order to find an algorithmic solution to Problem 3.18. To this end, this problem is restated as

$$\max_{\substack{\boldsymbol{Q}_{1,\ldots,K}\\ 1,\ldots,K\\ 1,\ldots,N}} \sum_{n=1}^{N}\sum_{j=1}^{K}\frac{1}{K}\log_2\left(\left|\boldsymbol{I}_t+\sum_{k=1}^{K}\boldsymbol{H}_k^{\mathrm{H}}\boldsymbol{Q}_{j,k,n}\boldsymbol{H}_k\right|\right),$$

subject to $\boldsymbol{Q}_{j,k,n} \geq \boldsymbol{0}$, $\forall j,k,n$, and $\sum_{n=1}^{N}\sum_{k=1}^{K}\mathrm{Tr}\left\{\boldsymbol{Q}_{[j+k]_K,k,n}\right\} \leq P$, $j=1,\ldots,K$. The equivalence between this problem and Problem 3.18 can be shown along the lines of the arguments provided in Section 3.1.1.1 to show the equivalence between Problem 3.7 and Problem 3.2. This formulation of the problem with decoupled matrix sets $\mathcal{S}_j = \{\boldsymbol{Q}_{[j+k]_K,k,n} | k=1,\ldots,K,\ n=1,\ldots,N\}$, $j=1,\ldots,K$, lends itself to a coordinate ascent algorithmic solution where at each step the objective function is increased by optimizing over the matrices of a certain set while keeping the matrices of all other sets fixed (see Algorithm 3.5). As in Algorithm 3.2, here also, due to concavity of the objective function, equality of all sets $\mathcal{S}_j$, $j=1,\ldots,K$, must hold at the final fixed point, i.e., $\boldsymbol{Q}_{1,k,n}=\cdots=\boldsymbol{Q}_{K,k,n}=\boldsymbol{Q}_{k,n}$, $\forall k,n$. As mentioned in Section 3.1.1.1, a drawback of this

Algorithm 3.5 OFDM sum-power iterative waterfilling (cyclic coordinate ascent)

1: $\boldsymbol{Q}_{j,k,n} \leftarrow \boldsymbol{0}, \quad k=1,\ldots,K,\ j=1,\ldots,K,\ n=1,\ldots,N, \quad i=1$
2: $R^{\mathrm{new}} \leftarrow 0$
3: **repeat**
4: $\quad R^{\mathrm{old}} \leftarrow R^{\mathrm{new}}$
5: $\quad$ **for** $n=1$ to N **do**
6: $\quad\quad$ **for** $m=1$ to K **do**
7: $\quad\quad\quad \boldsymbol{Z}_{[i+m]_K,n} \leftarrow \boldsymbol{I}_t + \sum_{k\neq m}\boldsymbol{H}_{k,n}^{\mathrm{H}}\boldsymbol{Q}_{[i+m]_K,k,n}\boldsymbol{H}_{k,n}$
8: $\quad\quad$ **end for**
9: $\quad$ **end for**
10: $\quad \mathcal{S}_i \leftarrow \underset{\substack{\boldsymbol{Q}_{1,\ldots,K}\\ 1,\ldots,N}}{\arg\max} \ \sum_{n=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\log_2\left(\left|\boldsymbol{Z}_{[i+k]_K,n}+\boldsymbol{H}_{k,n}^{\mathrm{H}}\boldsymbol{Q}_{k,n}\boldsymbol{H}_{k,n}\right|\right)$
$\quad$ subject to $\sum_{k=1,n=1}^{K,N}\mathrm{Tr}\{\boldsymbol{Q}_{k,n}\} \leq P$ and $\boldsymbol{Q}_{k,n}\geq\boldsymbol{0},\ \forall k,n$
11: $\quad i \leftarrow [i+1]_K$
12: $\quad R^{\mathrm{new}} \leftarrow \sum_{n=1}^{N}\frac{1}{K}\sum_{j=1}^{K}\log_2\left(\left|\boldsymbol{I}_t+\sum_{k=1}^{K}\boldsymbol{H}_{k,n}^{\mathrm{H}}\boldsymbol{Q}_{j,k,n}\boldsymbol{H}_{k,n}\right|\right)$
13: **until** $R^{\mathrm{new}}-R^{\mathrm{old}}<\epsilon$

kind of algorithms is storage capacity. In particular, this algorithm must keep a total of $NK(K-1)$ matrices in memory during execution. As we already saw, storage capacity can be reduced by introducing an averaging step at the end of each iteration, which, due to concavity, necessarily leads to an increase of the objective function. The resulting pseudocode is along the lines of Algorithm 3.3 and is given in Algorithm 3.6. This algorithm

has been previously reported in [115, 117]. Note that the complexity per iteration of both Algorithm 3.5 and Algorithm 3.6 is linear in the number of subcarriers. This is in contrast to the cubic complexity order resulting from a direct application of Algorithms 3.2 and 3.3 to the compact multiuser MIMO OFDM model in Eq. 3.13 without consideration of the underlying block diagonal structure.

Algorithm 3.6 OFDM sum-power iterative waterfilling

1: $\boldsymbol{Q}_{k,n}^{(0)} \leftarrow \boldsymbol{0}, \quad k = 1, \ldots, K,\ n = 1, \ldots, N,\ \ell = 0$
2: $R^{\text{new}} \leftarrow 0$
3: **repeat**
4: $\quad R^{\text{old}} \leftarrow R^{\text{new}}$
5: $\quad$ **for** $n = 1$ to N **do**
6: $\quad\quad$ **for** $m = 1$ to K **do**
7: $\quad\quad\quad \boldsymbol{Z}_{m,n} \leftarrow \boldsymbol{I}_t + \sum_{k \neq m} \boldsymbol{H}_{k,n}^{\text{H}} \boldsymbol{Q}_{k,n}^{(\ell)} \boldsymbol{H}_{k,n}$
8: $\quad\quad$ **end for**
9: $\quad$ **end for**
10: $\quad \boldsymbol{M}_{\substack{1,\ldots,K \\ 1,\ldots,N}} \leftarrow \underset{\substack{\boldsymbol{Q}_{1,\ldots,K} \\ 1,\ldots,N}}{\arg\max} \ \sum_{n=1}^{N} \frac{1}{K} \sum_{k=1}^{K} \log_2 \left(\left| \boldsymbol{Z}_{k,n} + \boldsymbol{H}_{k,n}^{\text{H}} \boldsymbol{Q}_{k,n} \boldsymbol{H}_{k,n} \right| \right)$

$\quad$ subject to $\sum_{k=1,n=1}^{K,N} \text{Tr}\{\boldsymbol{Q}_{k,n}\} \leq P$ and $\boldsymbol{Q}_{k,n} \geq \boldsymbol{0},\ \forall k, n$

11: $\quad \boldsymbol{Q}_{k,n}^{(\ell+1)} \leftarrow \frac{1}{K} \left(\boldsymbol{M}_{k,n} + (K-1) \boldsymbol{Q}_{k,n}^{(\ell)} \right),\ \forall k, n$
12: $\quad \ell \leftarrow \ell + 1$
13: $\quad R^{\text{new}} \leftarrow \sum_{n=1}^{N} \log_2 \left(\left| \boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_{k,n}^{\text{H}} \boldsymbol{Q}_{k,n}^{(\ell)} \boldsymbol{H}_{k,n} \right| \right)$
14: **until** $R^{\text{new}} - R^{\text{old}} < \epsilon$

3.1.2.2 Dual decomposition

In the previous section we have seen how the sum-power iterative waterfilling algorithm can be extended to a MIMO OFDM broadcast channel. In this section, we apply the dual decomposition technique to Problem 3.18. To this end, we rewrite Problem 3.18 as

$$\max_{\substack{\boldsymbol{Q}_{1,\ldots,K} \\ 1,\ldots,N}} \sum_{n=1}^{N} \log_2 \left| \boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_{k,n}^{\text{H}} \boldsymbol{Q}_{k,n} \boldsymbol{H}_{k,n} \right|, \tag{3.19}$$

subject to $\sum_{k=1}^{K} \text{Tr}\{\boldsymbol{Q}_{k,n}\} \leq P_n,\ \forall n,\ \boldsymbol{Q}_{k,n} \geq \boldsymbol{0},\ \forall k, n$ and $\sum_{n=1}^{N} P_n \leq P$. Here, $P_{1,\ldots,N}$ are auxiliary variables that represent the transmit power employed on each subcarrier. The Lagrange function of this problem with respect to the last constraint is given by

$$L \left(\boldsymbol{Q}_{\substack{1,\ldots,K \\ 1,\ldots,N}}, P_{1,\ldots,N}, \lambda \right) = \sum_{n=1}^{N} \log_2 \left(\left| \boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_{k,n}^{\text{H}} \boldsymbol{Q}_{k,n} \boldsymbol{H}_{k,n} \right| \right) + \lambda \left(P - \sum_{n=1}^{N} P_n \right),$$

and the corresponding Lagrangian dual function reads

$$g(\lambda) = \max_{\substack{\boldsymbol{Q}_{1,\ldots,K}, P_{1,\ldots,N} \\ 1,\ldots,N}} L\left(\boldsymbol{Q}_{\substack{1,\ldots,K \\ 1,\ldots,N}}, P_{1,\ldots,N}, \lambda\right), \tag{3.20}$$

subject to $\sum_{k=1}^{K} \mathrm{Tr}\{\boldsymbol{Q}_{k,n}\} \leq P_n$, $\forall n$, $\boldsymbol{Q}_{k,n} \geq \boldsymbol{0}$, $\forall k, n$. Since the primal problem is convex and strictly feasible arguments exist, strong duality holds. That is, the minimum of the dual problem

$$\min_{\lambda} \; g(\lambda),$$

subject to $\lambda \geq 0$, coincides with the maximum of Problem 3.19. Thus, an algorithmic approach can be followed that iteratively searches the minimum of the dual problem instead of trying to solve the primal problem directly.

Assume $\lambda^{(\ell)}$ to be the value obtained for λ after the ℓth iteration. Computation of $g(\lambda^{(\ell)})$ is done by solving Problem 3.20. By noting that

$$\begin{aligned}\max_{\substack{\boldsymbol{Q}_{1,\ldots,K}, P_{1,\ldots,N} \\ 1,\ldots,N}} & L\left(\boldsymbol{Q}_{\substack{1,\ldots,K \\ 1,\ldots,N}}, P_{1,\ldots,N}, \lambda^{(\ell)}\right) = \\ &= P + \sum_{n=1}^{N} \max_{\boldsymbol{Q}_{1,\ldots,K,n}, P_n} \left(\log_2\left(\left|\boldsymbol{I}_t + \sum_{k=1}^{K} \boldsymbol{H}_{k,n}^{\mathrm{H}} \boldsymbol{Q}_{k,n} \boldsymbol{H}_{k,n}\right|\right) - \lambda^{(\ell)} P_n\right),\end{aligned}$$

subject to $\sum_{k=1}^{K} \mathrm{Tr}\{\boldsymbol{Q}_{k,n}\} \leq P_n$, $\forall n$, $\boldsymbol{Q}_{k,n} \geq \boldsymbol{0}$, $\forall k, n$, we observe that this problem decomposes into N separate maximization problems, each problem corresponding to a different subcarrier. For any subcarrier n, the KKT conditions of the corresponding subproblem yield

$$\hat{\boldsymbol{H}}_{j,n}\left(\boldsymbol{I}_t + \hat{\boldsymbol{H}}_{j,n}^{\mathrm{H}} \boldsymbol{Q}_j \hat{\boldsymbol{H}}_{j,n}\right)^{-1} \hat{\boldsymbol{H}}_{j,n}^{\mathrm{H}} + \boldsymbol{\Phi}_{j,n} - \lambda^{(\ell)} \boldsymbol{I}_{r_j} = \boldsymbol{0}, \quad \forall j \tag{3.21}$$

where

$$\hat{\boldsymbol{H}}_{j,n} = \boldsymbol{H}_{j,n}\left(\boldsymbol{I}_t + \sum_{k \neq j} \boldsymbol{H}_{k,n}^{\mathrm{H}} \boldsymbol{Q}_{k,n} \boldsymbol{H}_{k,n}\right)^{-1/2}$$

is the effective channel seen by user j at subcarrier n and $\boldsymbol{\Phi}_{j,n} \geq \boldsymbol{0}$ is the Lagrange multiplier associated with the constraint $\boldsymbol{Q}_{j,n} \geq \boldsymbol{0}$. Note that Eq. 3.21 has the form of the optimality conditions given in Eqs. 3.12 and 3.6. Therefore, we can proceed, as described in Section 3.1.1.2, by using Algorithm 3.1 in order to find a solution to each of the N subproblems, which immediately leads to a solution for the composite Problem 3.20. Note that here, the same as in Section 3.1.1.2, the power constraint is implicitly replaced by the multiplier $\lambda^{(\ell)}$, which determines the water level and is constant across users and subcarriers.

Once $g(\lambda^{(\ell)})$ has been evaluated, a subgradient of this function at $\lambda^{(\ell)}$ must be found, which gives an appropriate direction for the computation of $\lambda^{(\ell+1)}$. Assume that $\boldsymbol{Q}_{k,n}^{(\ell)}$, $k = 1, \ldots, K$, $n = 1, \ldots, N$, and $P_{1,\ldots,N}^{(\ell)}$ are maximizers of 3.20. From the definition of

subgradient in Eq. A.11 and the inequalities

$$\begin{aligned} g(\lambda^{(\ell)}+\Delta\lambda) &- g(\lambda^{(\ell)}) = \\ &= \max_{\substack{\boldsymbol{Q}_{1,\ldots,K}, P_{1,\ldots,N} \\ 1,\ldots,N}} L\left(\underset{1,\ldots,N}{\boldsymbol{Q}_{1,\ldots,K}}, P_{1,\ldots,N}, \lambda^{(\ell)}+\Delta\lambda\right) - L\left(\underset{1,\ldots,N}{\boldsymbol{Q}^{(\ell)}_{1,\ldots,K}}, P^{(\ell)}_{1,\ldots,N}, \lambda^{(\ell)}\right) \\ &\geq L\left(\underset{1,\ldots,N}{\boldsymbol{Q}^{(\ell)}_{1,\ldots,K}}, P^{(\ell)}_{1,\ldots,N}, \lambda^{(\ell)}+\Delta\lambda\right) - L\left(\underset{1,\ldots,N}{\boldsymbol{Q}^{(\ell)}_{1,\ldots,K}}, P^{(\ell)}_{1,\ldots,N}, \lambda^{(\ell)}\right) \\ &= \Delta\lambda\left(P - \sum_{n=1}^{N} P_n^{(\ell)}\right), \end{aligned}$$

it becomes clear that

$$s^{(\ell)} = \left(P - \sum_{n=1}^{N} P_n^{(\ell)}\right)$$

is a subgradient of $g(\lambda)$ at $\lambda^{(\ell)}$. A convenient search direction is given by $-s^{(\ell)}$, i.e., if the constraint is violated $\lambda^{(\ell)}$ should be increased. Otherwise, $\lambda^{(\ell)}$ should be decreased. Here, as in Section 3.1.1.2, due to the monotonicity of the constraint with respect to λ, bisection can be used in order to choose the step size of each update. As shown in Section 3.1.1.2 and Appendix A.3, given the optimum multiplier λ, the maximizers in the definition of the Lagrangian dual function, i.e., Problem 3.20, turn out to be the maximizers of the primal, i.e., Problem 3.19. This property can be used in order to compute the optimum covariance matrices once the optimum λ has been found. The pseudocode of this algorithm is given in Algorithm 3.7. Note that the complexity per iteration of this algorithm is linear in the number of subcarriers. Indeed, each iteration essentially has the complexity of N sum-rate maximization problems for which the input parameters do not depend on N. A similar approach has been followed in [68] in order to extend a weighted sum-rate maximizing algorithm for memoryless broadcast channels with single receive antennas [69] to a multicarrier setting.

In these last two sections we have extended the, probably, two most significant algorithms for sum-rate maximization in memoryless MIMO broadcast channels to a MIMO OFDM broadcast setting by applying a cyclic expansion technique of the original problem in Section 3.1.2.1 and using dual decomposition in this section. An alternative and general way to extend sum-rate maximizing algorithms for memoryless channels to time-dispersive OFDM channels was presented in [83] and is based on a factorization of the input covariance matrices. This technique will be used in the next chapter to extend weighted sum-rate maximization algorithms for memoryless channels to time-dispersive channels.

3.2 Weighted sum rate

While for specific scenarios sum-rate maximization might be satisfactory, in most scenarios this criterion results in an uneven distribution of rates that might be unwished. For instance, it might happen that, for a given network, some users with weak channels are not served at all and, at the same time, some other users obtain far too high rates for the

Algorithm 3.7 Sum-rate maximization in MIMO OFDM based on dual decomposition

Initialize $\lambda_{\min}$ and $\lambda_{\max}$
$\lambda^{(0)} = (\lambda_{\max} + \lambda_{\min})/2$, $\ell \leftarrow 0$
repeat
  Compute $g(\lambda^{(\ell)})$ by applying Algorithm 3.1 to each subcarrier
  if $\left(P - \sum_{n=1}^{N} P_n^{(\ell)}\right) > 0$ **then**
    $\lambda^{(\ell+1)} \leftarrow \frac{\lambda^{(\ell)}+\lambda_{\min}}{2}$
    $\lambda_{\max} \leftarrow \lambda^{(\ell)}$
  else
    $\lambda^{(\ell+1)} \leftarrow \frac{\lambda^{(\ell)}+\lambda_{\max}}{2}$
    $\lambda_{\min} \leftarrow \lambda^{(\ell)}$
  end if
  $\ell \leftarrow \ell + 1$
until $\lambda_{\max} - \lambda_{\min} < \epsilon$

service that they requested. In order to prevent such outcomes, it is advisable to choose transmission parameters according to criteria that include some mechanism to control the final distribution of resources among users. A first approach that allows some control on the quality of service finally obtained by the users in the network is weighted sum-rate maximization. In this approach, the rates of the users are weighted with so-called priorities, which, as the name indicates, establish a ranking among users according to the quality of service that they should be provided with.

Assume a Gaussian broadcast channel as given by Eq. 2.6, a DPC-based successive encoding scheme and let π be a permutation function defining the order in which information for the different users in the network is encoded. Mathematically, we can state the weighted sum-rate maximization problem as follows,

$$\max_{\pi, \boldsymbol{\Sigma}_{1,\ldots,K}} \sum_{k=1}^{K} \mu_k R_k \left(\pi, \boldsymbol{\Sigma}_{1,\ldots,K}\right),$$

subject to $\sum_{k=1}^{K} \mathrm{Tr}\left\{\boldsymbol{\Sigma}_k\right\} \leq P$ and $\boldsymbol{\Sigma}_k \geq \mathbf{0}$, $\forall k$. Here $\mu_k \in \mathbb{R}_+$, $k = 1, \ldots, K$, are the priorities or weights and, for any user $\pi(k')$, the rate $R_\pi(k')$ is given by Eq. 3.1. As we have already mentioned in the discussion of the sum-rate maximization problem, for a fixed π, $R_k\left(\pi, \boldsymbol{\Sigma}_{1,\ldots,K}\right)$ is, in general, neither a concave nor a convex function of the matrices $\boldsymbol{\Sigma}_{1,\ldots,K}$. As a consequence, the same holds for the weighted sum of these rates. That is, the same as the sum-rate maximization problem, the weighted sum-rate maximization problem stated in this way does not qualify as a convex optimization problem either. Fortunately, also in this case, we can resort to duality in order to restate the original problem so that convexity eventually holds. Doing so, the following equivalent weighted sum-rate maximization problem can be written,

$$\max_{\bar{\pi}, \boldsymbol{Q}_{1,\ldots,K}} \sum_{k=1}^{K} \mu_k R_k \left(\bar{\pi}, \boldsymbol{Q}_{1,\ldots,K}\right), \tag{3.22}$$

subject to $\sum_{k=1}^{K} \mathrm{Tr}\{\boldsymbol{Q}_k\} \leq P$ and $\boldsymbol{Q}_k \geq \boldsymbol{0}$, $\forall k$. Here, $\bar{\pi}$ is a permutation function indicating the order in which users are decoded and, for any user $\bar{\pi}(k')$, the corresponding rate is given by Eq. 3.3. From the fact that $\mathcal{R}^{\mathrm{MAC}}(\boldsymbol{Q}_{1,\ldots,K})$ is a polymatroid and the polymatroid characterization in terms of a weighted sum-rate maximization problem given in [121, Lemma 3.2], it immediately follows that for any set of covariance matrices $\boldsymbol{Q}_{1,\ldots,K}$ the optimum decoding order is such that $\mu_{\bar{\pi}(K)} \geq \mu_{\bar{\pi}(K-1)} \geq \cdots \geq \mu_{\bar{\pi}(1)}$. That is, optimally, users with higher priority are decoded last and users with lower priority first. Considering this optimum order and substituting Eq. 3.3 in Problem 3.22 we obtain the following formulation for the weighted sum-rate maximization problem

$$\max_{\boldsymbol{Q}_{1,\ldots,K}} \sum_{k=1}^{K} \eta_{\bar{\pi}(k)} \log_2 \left(\left| \boldsymbol{I}_t + \sum_{j \geq k} \boldsymbol{H}_{\bar{\pi}(j)}^{\mathrm{H}} \boldsymbol{Q}_{\bar{\pi}(j)} \boldsymbol{H}_{\bar{\pi}(j)} \right| \right), \tag{3.23}$$

subject to $\sum_{k=1}^{K} \mathrm{Tr}\{\boldsymbol{Q}_k\} \leq P$ and $\boldsymbol{Q}_k \geq \boldsymbol{0}$, $\forall k$, where $\eta_{\bar{\pi}(k)} = \mu_{\bar{\pi}(k)} - \mu_{\bar{\pi}(k-1)}$, for $k = 2, \ldots, K$, and $\eta_{\bar{\pi}(1)} = \mu_{\bar{\pi}(1)}$. Observe that $\eta_k \geq 0$, $\forall k$. Thus, the objective function of Problem 3.23 is obtained by adding K concave functions of the input covariance matrices and, therefore, it is concave. Since the feasible set is convex, the global optimum can be achieved by employing iterative local search algorithms. Before we describe some of the most significant algorithmical approaches to solve Problem 3.23 in the next section, let us conclude this section by discussing the relationship between the weighted sum-rate maximization problem and the geometry of the capacity region of the Gaussian MIMO broadcast channel.

In the last chapter we saw that the capacity region of the Gaussian MIMO broadcast channel is equal to the capacity region of the Gaussian MIMO multiple access channel $\mathcal{R}^{\mathrm{MAC}}(P)$. This region is defined as the union of polymatroids that result from all possible transmit statistics satisfying a transmit power constraint jointly imposed on all users of the network (cf. Eq. 2.29). Thus, any point on the boundary of $\mathcal{R}^{\mathrm{MAC}}(P)$ is either a vertex or lies on an edge or face of a particular polymatroid corresponding to certain statistics. While points on the boundary of $\mathcal{R}^{\mathrm{MAC}}(P)$ corresponding to vertices are obtained as solutions to Problem 3.22, points on the boundary of $\mathcal{R}^{\mathrm{MAC}}(P)$ corresponding to edges or faces of polymatroids can not be characterized as solutions to this problem. This is due to the fact that in the formulation of Problem 3.22 we have implicitly restricted ourselves to those rate vectors that are achievable by successive decoding (cf. Section 2.2.2.3).

Consider Problem 3.23 with the optimum decoding order for the weights $\mu_1 \leq \cdots \leq \mu_{\bar{k}} \leq \mu_{\bar{k}+1} = \cdots = \mu_{\bar{k}+J} \leq \cdots \leq \mu_K$. Rewriting the objective function of this problem using priorities $\mu_{1,\ldots,K}$ rather than $\eta_{1,\ldots,K}$ we obtain

$$\begin{aligned} &\sum_{k=1}^{\bar{k}} \mu_k \log_2 \left(\frac{|\boldsymbol{I}_t + \sum_{j \geq k} \boldsymbol{H}_j^{\mathrm{H}} \boldsymbol{Q}_j \boldsymbol{H}_j|}{|\boldsymbol{I}_t + \sum_{j > k} \boldsymbol{H}_j^{\mathrm{H}} \boldsymbol{Q}_j \boldsymbol{H}_j|} \right) + \mu \log_2 \left(\frac{|\boldsymbol{I}_t + \sum_{j \geq \bar{k}+1} \boldsymbol{H}_j^{\mathrm{H}} \boldsymbol{Q}_j \boldsymbol{H}_j|}{|\boldsymbol{I}_t + \sum_{j \geq \bar{k}+J+1} \boldsymbol{H}_j^{\mathrm{H}} \boldsymbol{Q}_j \boldsymbol{H}_j|} \right) + \\ &\quad + \sum_{k=\bar{k}+J+1}^{K} \mu_k \log_2 \left(\frac{|\boldsymbol{I}_t + \sum_{j \geq k} \boldsymbol{H}_j^{\mathrm{H}} \boldsymbol{Q}_j \boldsymbol{H}_j|}{|\boldsymbol{I}_t + \sum_{j > k} \boldsymbol{H}_j^{\mathrm{H}} \boldsymbol{Q}_j \boldsymbol{H}_j|} \right), \end{aligned}$$

where $\mu = \mu_{\bar{k}+1} = \cdots = \mu_{\bar{k}+J}$. Observe that this function does not depend on the individual rates of users $\bar{k}+1, \ldots, \bar{k}+J$, but only on the sum rate achieved by these users, which is

given by the second term in the above expression. Due to strict concavity of the objective function, there is a unique set of covariance matrices $\boldsymbol{Q}_{1,\ldots,K}$ that achieve the optimum value of this function. However, there are at least $J!$ rate vectors that can be achieved by performing successive decoding and are optimum.[1] Each of these vectors corresponds to a different decoding order within the group of users $\bar{k}+1,\ldots,\bar{k}+J$. In terms of geometry, we can say that, in this case, the polymatroid $\mathcal{R}^{\mathrm{MAC}}(\boldsymbol{Q}_{1,\ldots,K})$ defined by the optimum covariance matrices has a face with $J!$ vertices whose points all represent optimum rate vectors for the weighted sum-rate maximization problem. All points on this face lie on the boundary of $\mathcal{R}^{\mathrm{MAC}}(P)$ as points lying in the interior of $\mathcal{R}^{\mathrm{MAC}}(P)$ can never be weighted sum-rate maximizers. As it has been already discussed (cf. Section 2.2.2.3), the vertices are achieved by successive decoding in the MAC or DPC-based successive encoding in the BC. All other points on this face can be reached by time sharing. Hence, we observe that weighted sum-rate maximization with equality in the priorities of certain users give rise to statistics that are associated with time-sharing regions on the boundary of $\mathcal{R}^{\mathrm{MAC}}(P)$ or, equivalently, $\mathcal{R}^{\mathrm{DPC}}(P)$. In the previous chapter, Conjecture 2.2.3 states that all points on the dominant face of the sum-rate maximizing polymatroid, i.e., all weights equal, might be achievable in the BC without resorting to time-sharing. Here, Conjecture 3.2.1 extends this hypothesis to all other time-sharing regions on the boundary of the capacity region.

Conjecture 3.2.1. *Given a Gaussian broadcast channel, let $\boldsymbol{Q}_{1,\ldots,K}$ be the set of covariance matrices that maximize $\sum_{k=1}^{K}\mu_k R_k$ with $\mu_1 \leq \cdots \leq \mu_{\bar{k}} \leq \mu_{\bar{k}+1} = \cdots = \mu_{\bar{k}+J} \leq \cdots \leq \mu_K$ for a given transmit power constraint in the dual MAC. Additionally, let $\boldsymbol{\Sigma}_{\bar{k}+1,\ldots,\bar{k}+J}$ be a set of transmit covariance matrices achieving any of the vertices of the polymatroid $\mathcal{R}^{\mathrm{MAC}}\left(\boldsymbol{Q}_{\bar{k}+1,\ldots,\bar{k}+J}\right)$ in the BC by encoding users $\bar{k}+1,\ldots,\bar{k}+J$ successively with order π according to the DPC scheme. Then,*

$$\mathcal{R}^{\mathrm{Marton}}\left(\pi, \boldsymbol{\Sigma}_{\bar{k}+1,\ldots,\bar{k}+J}\right) = \mathcal{R}^{\mathrm{MAC}}\left(\boldsymbol{Q}_{\bar{k}+1,\ldots,\bar{k}+J}\right).$$

In the following, the validity of this conjecture will be shown for $K=3$, $J=2$ and $r_k=1$. Consider the BC given by

$$y_k = \boldsymbol{h}_k^{\mathrm{H}}\boldsymbol{x} + n_k, \quad k=1,2,3.$$

Let us start assuming $\mu_1=\mu_2\leq\mu_3$. The objective function for the weighted sum-rate maximization problem in the dual MAC is given by

$$\begin{aligned}\sum_{k=1}^{3}\mu_k R_k =& \mu_1 \log_2\left(\left|\boldsymbol{I}+\boldsymbol{h}_1\boldsymbol{h}_1^{\mathrm{H}}q_1+\boldsymbol{h}_2\boldsymbol{h}_2^{\mathrm{H}}q_2+\boldsymbol{h}_3\boldsymbol{h}_3^{\mathrm{H}}q_3\right|\right)\\ &-\mu_1\log_2\left(\left|\boldsymbol{I}+\boldsymbol{h}_3\boldsymbol{h}_3^{\mathrm{H}}q_3\right|\right)+\mu_3\log_2\left(\left|\boldsymbol{I}+\boldsymbol{h}_3\boldsymbol{h}_3^{\mathrm{H}}q_3\right|\right). \qquad (3.24)\end{aligned}$$

Note that users 1 and 2 are decoded before user 3. That is, the signal transmitted by user 3 is viewed as interference while detecting the signals from users 1 and 2. Define the effective channels of these users as

$$\hat{\boldsymbol{h}}_k = \left(\boldsymbol{I}+\boldsymbol{h}_3\boldsymbol{h}_3^{\mathrm{H}}q_3\right)^{-1}\boldsymbol{h}_k, \quad k=1,2.$$

[1] This number becomes larger if equality also holds for the weights of some other users different from $\bar{k}+1,\ldots,\bar{k}+J$.

Using this definition Eq. 3.24 can be written as

$$\sum_{k=1}^{3} \mu_k R_k = \mu_1 \log_2 \left(\left| \boldsymbol{I} + \hat{\boldsymbol{h}}_1 \hat{\boldsymbol{h}}_1^{\mathrm{H}} q_1 + \hat{\boldsymbol{h}}_2 \hat{\boldsymbol{h}}_2^{\mathrm{H}} q_2 \right| \right) + \mu_3 \log_2 \left(\left| \boldsymbol{I} + \boldsymbol{h}_3 \boldsymbol{h}_3^{\mathrm{H}} q_3 \right| \right).$$

Considering the transmit power constraint $q_1 + q_2 + q_3 \leq P$, the KKT conditions for optimality of q_1 and q_2 yield

$$\hat{\boldsymbol{h}}_1^{\mathrm{H}} \left(\boldsymbol{I} + \hat{\boldsymbol{h}}_1 \hat{\boldsymbol{h}}_1^{\mathrm{H}} q_1 + \hat{\boldsymbol{h}}_2 \hat{\boldsymbol{h}}_2^{\mathrm{H}} q_2 \right)^{-1} \hat{\boldsymbol{h}}_1 = \hat{\boldsymbol{h}}_2^{\mathrm{H}} \left(\boldsymbol{I} + \hat{\boldsymbol{h}}_1 \hat{\boldsymbol{h}}_1^{\mathrm{H}} q_1 + \hat{\boldsymbol{h}}_2 \hat{\boldsymbol{h}}_2^{\mathrm{H}} q_2 \right)^{-1} \hat{\boldsymbol{h}}_2.$$

This is exactly the result given by Eq. 2.30 in Section 2.2.2.3 as necessary condition for sum-rate optimality. That is, optimality requires that users 1 and 2 achieve maximum sum rate with the power left by user 3. As we already saw, a consequence of this condition is $\mathcal{R}^{\mathrm{Marton}}(\pi, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) = \mathcal{R}^{\mathrm{MAC}}(q_1, q_2)$.

Now, let us assume $\mu_1 \leq \mu_2 = \mu_3$. Different to the previous case, the two users with equal weights do not suffer interference from the other user. Rather, these two users cause interference to the third one and, therefore, sum-rate maximization of these users is not necessarily optimum. That is, the optimization of q_2 and q_3 can not be reduced to a sum-rate maximization problem, which calls for a more elaborated proof. The objective function is now given by

$$\sum_{k=1}^{3} \mu_k R_k = \mu_1 \log_2 \left(\left| \boldsymbol{I} + \boldsymbol{h}_1 \boldsymbol{h}_1^{\mathrm{H}} q_1 + \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}} q_2 + \boldsymbol{h}_3 \boldsymbol{h}_3^{\mathrm{H}} q_3 \right| \right) + \eta \log_2 \left(\left| \boldsymbol{I} + \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}} q_2 + \boldsymbol{h}_3 \boldsymbol{h}_3^{\mathrm{H}} q_3 \right| \right).$$

where $\eta = \mu_2 - \mu_1$. In this case, the KKT optimality conditions yield

$$\begin{aligned}
\lambda &= \mu_1 \boldsymbol{h}_1^{\mathrm{H}} \left(\boldsymbol{I} + \boldsymbol{h}_1 \boldsymbol{h}_1^{\mathrm{H}} q_1 + \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}} q_2 + \boldsymbol{h}_3 \boldsymbol{h}_3^{\mathrm{H}} q_3 \right)^{-1} \boldsymbol{h}_1, \\
\lambda &= \boldsymbol{h}_2^{\mathrm{H}} \left(\mu_1 \left(\boldsymbol{I} + \boldsymbol{h}_1 \boldsymbol{h}_1^{\mathrm{H}} q_1 + \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}} q_2 + \boldsymbol{h}_3 \boldsymbol{h}_3^{\mathrm{H}} q_3 \right)^{-1} + \eta \left(\boldsymbol{I} + \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}} q_2 + \boldsymbol{h}_3 \boldsymbol{h}_3^{\mathrm{H}} q_3 \right)^{-1} \right) \boldsymbol{h}_2, \\
\lambda &= \boldsymbol{h}_3^{\mathrm{H}} \left(\mu_1 \left(\boldsymbol{I} + \boldsymbol{h}_1 \boldsymbol{h}_1^{\mathrm{H}} q_1 + \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}} q_2 + \boldsymbol{h}_3 \boldsymbol{h}_3^{\mathrm{H}} q_3 \right)^{-1} + \eta \left(\boldsymbol{I} + \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}} q_2 + \boldsymbol{h}_3 \boldsymbol{h}_3^{\mathrm{H}} q_3 \right)^{-1} \right) \boldsymbol{h}_3,
\end{aligned}$$

where λ is the Lagrangian multiplier associated with the transmit power constraint. To be strict, these are the optimality conditions under the assumption that all users receive some power. If switching off user 1 turns out to be optimum, the problem degenerates into a sum-rate maximization problem for users 2 and 3 for which we know that $\mathcal{R}^{\mathrm{Marton}}(\pi, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3) = \mathcal{R}^{\mathrm{MAC}}(q_2, q_3)$ obtains. Coming back to the general case, in which all users are on, optimally, user 1 is decoded first, while the decoding order of users 2 and 3 can be arbitrarily chosen. Let us assume that user 2 is the second decoded user and user 3 the third. Under this decoding order, the transmit covariance matrices in the BC channel corresponding to the optimum powers q_1, q_2 and q_3 in the dual MAC can be computed as [128]

$$\boldsymbol{\Sigma}_1 = \frac{\left(\boldsymbol{I} + \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}} q_2 + \boldsymbol{h}_3 \boldsymbol{h}_3^{\mathrm{H}} q_3 \right)^{-1} \boldsymbol{h}_1 q_1 \boldsymbol{h}_1^{\mathrm{H}} \left(\boldsymbol{I} + \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}} q_2 + \boldsymbol{h}_3 \boldsymbol{h}_3^{\mathrm{H}} q_3 \right)^{-1}}{\boldsymbol{h}_1^{\mathrm{H}} \left(\boldsymbol{I} + \boldsymbol{h}_2 \boldsymbol{h}_2^{\mathrm{H}} q_2 + \boldsymbol{h}_3 \boldsymbol{h}_3^{\mathrm{H}} q_3 \right)^{-1} \boldsymbol{h}_1}, \tag{3.25}$$

$$\boldsymbol{\Sigma}_2 = \frac{\left(\boldsymbol{I} + \boldsymbol{h}_3 \boldsymbol{h}_3^{\mathrm{H}} q_3 \right)^{-1} \boldsymbol{h}_2 q_2 \boldsymbol{h}_2^{\mathrm{H}} \left(\boldsymbol{I} + \boldsymbol{h}_3 \boldsymbol{h}_3^{\mathrm{H}} q_3 \right)^{-1}}{\boldsymbol{h}_2^{\mathrm{H}} \left(\boldsymbol{I} + \boldsymbol{h}_3 \boldsymbol{h}_3^{\mathrm{H}} q_3 \right)^{-1} \boldsymbol{h}_2} \left(1 + \boldsymbol{h}_2^{\mathrm{H}} \boldsymbol{\Sigma}_1 \boldsymbol{h}_2 \right), \tag{3.26}$$

$$\boldsymbol{\Sigma}_3 = \frac{\boldsymbol{h}_3 q_3 \boldsymbol{h}_3^{\mathrm{H}}}{\boldsymbol{h}_3^{\mathrm{H}} \boldsymbol{h}_3} \left(1 + \boldsymbol{h}_3^{\mathrm{H}} \boldsymbol{\Sigma}_1 \boldsymbol{h}_3 + \boldsymbol{h}_3^{\mathrm{H}} \boldsymbol{\Sigma}_2 \boldsymbol{h}_3 \right). \tag{3.27}$$

In order to prove that $\mathcal{R}^{\text{Marton}}(\pi, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3) = \mathcal{R}^{\text{MAC}}(q_2, q_3)$, with $\pi(1) = 3$ and $\pi(2) = 2$, we show that the rate increase experienced by user 2 when reversing π and maintaining the statistics in the broadcast channel is the same as the rate increase experienced by this user in the dual MAC if it is decoded after user 3. Recalling the discussion in Section 2.2.2.3 concerning the proof of conjecture 2.2.3 for a particular case, this is equivalent to showing

$$\log_2\left(|\boldsymbol{I}_t + q_2\boldsymbol{h}_2\boldsymbol{h}_2^{\text{H}}||\boldsymbol{I}_t + q_3\boldsymbol{h}_3\boldsymbol{h}_3^{\text{H}}||\boldsymbol{I}_t + q_2\boldsymbol{h}_2\boldsymbol{h}_2^{\text{H}} + q_3\boldsymbol{h}_3\boldsymbol{h}_3^{\text{H}}|^{-1}\right) = \\ = \log_2\left(1 + \frac{\boldsymbol{h}_2^{\text{H}}\boldsymbol{\Sigma}_2\boldsymbol{h}_2\boldsymbol{h}_2^{\text{H}}\boldsymbol{\Sigma}_3\boldsymbol{h}_2}{\left(1 + \boldsymbol{h}_2^{\text{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_2 + \boldsymbol{h}_2^{\text{H}}\boldsymbol{\Sigma}_2\boldsymbol{h}_2\right)^2}\right). \quad (3.28)$$

Note that, apart from the indexes, the only change in this expression with respect to Eq. 2.36 is the inclusion in the denominator on the right-hand side of an interference term caused by user 1.

Applying the matrix inversion lemma to the factors $\left(\boldsymbol{I} + \boldsymbol{h}_1\boldsymbol{h}_1^{\text{H}}q_1 + \boldsymbol{h}_2\boldsymbol{h}_2^{\text{H}}q_2 + \boldsymbol{h}_3\boldsymbol{h}_3^{\text{H}}q_3\right)^{-1}$, the optimality conditions given above can be rewritten as

$$\begin{aligned}
\lambda &= \mu_1\frac{\boldsymbol{h}_1^{\text{H}}\left(\boldsymbol{I} + \boldsymbol{h}_2\boldsymbol{h}_2^{\text{H}}q_2 + \boldsymbol{h}_3\boldsymbol{h}_3^{\text{H}}q_3\right)^{-1}\boldsymbol{h}_1}{1 + q_1\boldsymbol{h}_1^{\text{H}}\left(\boldsymbol{I} + \boldsymbol{h}_2\boldsymbol{h}_2^{\text{H}}q_2 + \boldsymbol{h}_3\boldsymbol{h}_3^{\text{H}}q_3\right)^{-1}\boldsymbol{h}_1}, \\
\lambda &= -\mu_1\frac{q_1\left|\boldsymbol{h}_2^{\text{H}}\left(\boldsymbol{I} + \boldsymbol{h}_2\boldsymbol{h}_2^{\text{H}}q_2 + \boldsymbol{h}_3\boldsymbol{h}_3^{\text{H}}q_3\right)^{-1}\boldsymbol{h}_1\right|^2}{1 + q_1\boldsymbol{h}_1^{\text{H}}\left(\boldsymbol{I} + \boldsymbol{h}_2\boldsymbol{h}_2^{\text{H}}q_2 + \boldsymbol{h}_3\boldsymbol{h}_3^{\text{H}}q_3\right)^{-1}\boldsymbol{h}_1} + \mu_2\boldsymbol{h}_2^{\text{H}}\left(\boldsymbol{I} + \boldsymbol{h}_2\boldsymbol{h}_2^{\text{H}}q_2 + \boldsymbol{h}_3\boldsymbol{h}_3^{\text{H}}q_3\right)^{-1}\boldsymbol{h}_2, \\
\lambda &= -\mu_1\frac{q_1\left|\boldsymbol{h}_3^{\text{H}}\left(\boldsymbol{I} + \boldsymbol{h}_2\boldsymbol{h}_2^{\text{H}}q_2 + \boldsymbol{h}_3\boldsymbol{h}_3^{\text{H}}q_3\right)^{-1}\boldsymbol{h}_1\right|^2}{1 + q_1\boldsymbol{h}_1^{\text{H}}\left(\boldsymbol{I} + \boldsymbol{h}_2\boldsymbol{h}_2^{\text{H}}q_2 + \boldsymbol{h}_3\boldsymbol{h}_3^{\text{H}}q_3\right)^{-1}\boldsymbol{h}_1} + \mu_2\boldsymbol{h}_3^{\text{H}}\left(\boldsymbol{I} + \boldsymbol{h}_2\boldsymbol{h}_2^{\text{H}}q_2 + \boldsymbol{h}_3\boldsymbol{h}_3^{\text{H}}q_3\right)^{-1}\boldsymbol{h}_3.
\end{aligned}$$

From Eq. 3.25 and the two first conditions we obtain

$$1 + \boldsymbol{h}_2^{\text{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_2 = \mu\boldsymbol{h}_2^{\text{H}}\left(\boldsymbol{I} + \boldsymbol{h}_2\boldsymbol{h}_2^{\text{H}}q_2 + \boldsymbol{h}_3\boldsymbol{h}_3^{\text{H}}q_3\right)^{-1}\boldsymbol{h}_2, \quad (3.29)$$

where $\mu = \mu_2/\lambda$. Similarly, from the first and the third condition we get

$$1 + \boldsymbol{h}_3^{\text{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_3 = \mu\boldsymbol{h}_3^{\text{H}}\left(\boldsymbol{I} + \boldsymbol{h}_2\boldsymbol{h}_2^{\text{H}}q_2 + \boldsymbol{h}_3\boldsymbol{h}_3^{\text{H}}q_3\right)^{-1}\boldsymbol{h}_3. \quad (3.30)$$

Using these expressions we shall first simplify the right-hand side of Eq. 3.28. From substitution of Eq. 3.26 in the right-hand side of Eq. 3.28 we have

$$\begin{aligned}
&\log_2\left(1 + \frac{\boldsymbol{h}_2^{\text{H}}\boldsymbol{\Sigma}_2\boldsymbol{h}_2\boldsymbol{h}_2^{\text{H}}\boldsymbol{\Sigma}_3\boldsymbol{h}_2}{\left(1 + \boldsymbol{h}_2^{\text{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_2 + \boldsymbol{h}_2^{\text{H}}\boldsymbol{\Sigma}_2\boldsymbol{h}_2\right)^2}\right) = \\
&\quad = \log_2\left(1 + \frac{q_2\boldsymbol{h}_2^{\text{H}}\left(\boldsymbol{I} + \boldsymbol{h}_3\boldsymbol{h}_3^{\text{H}}q_3\right)^{-1}\boldsymbol{h}_2}{1 + q_2\boldsymbol{h}_2^{\text{H}}\left(\boldsymbol{I} + \boldsymbol{h}_3\boldsymbol{h}_3^{\text{H}}q_3\right)^{-1}\boldsymbol{h}_2}\frac{\boldsymbol{h}_2^{\text{H}}\boldsymbol{\Sigma}_3\boldsymbol{h}_2}{\left(1 + \boldsymbol{h}_2^{\text{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_2 + \boldsymbol{h}_2^{\text{H}}\boldsymbol{\Sigma}_2\boldsymbol{h}_2\right)}\right) \\
&\quad = \log_2\left(1 + \boldsymbol{h}_2^{\text{H}}\left(\boldsymbol{I} + \boldsymbol{h}_2^{\text{H}}\boldsymbol{h}_2q_2 + \boldsymbol{h}_3\boldsymbol{h}_3^{\text{H}}q_3\right)^{-1}\boldsymbol{h}_2\frac{\boldsymbol{h}_2^{\text{H}}\boldsymbol{\Sigma}_3\boldsymbol{h}_2}{\left(1 + \boldsymbol{h}_2^{\text{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_2 + \boldsymbol{h}_2^{\text{H}}\boldsymbol{\Sigma}_2\boldsymbol{h}_2\right)}\right) \quad (3.31)
\end{aligned}$$

where the matrix inversion lemma has been applied in order to obtain the second equality. From Eqs. 3.26, 3.27, 3.29, 3.30 and application of the matrix inversion lemma we compute

$$\boldsymbol{h}_2^{\mathrm{H}}\boldsymbol{\Sigma}_3\boldsymbol{h}_2 = \mu q_3 \frac{|\boldsymbol{h}_2^{\mathrm{H}}\boldsymbol{h}_3|^2}{\|\boldsymbol{h}_3\|_2^2}\boldsymbol{h}_3^{\mathrm{H}}\left(\boldsymbol{I}+\boldsymbol{h}_3\boldsymbol{h}_3^{\mathrm{H}}q_3\right)^{-1}\boldsymbol{h}_3. \tag{3.32}$$

Using Eqs. 3.26 and 3.29 we compute

$$\begin{aligned}1+\boldsymbol{h}_2^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_2+&\boldsymbol{h}_2^{\mathrm{H}}\boldsymbol{\Sigma}_2\boldsymbol{h}_2 = \\ &=\mu\boldsymbol{h}_2^{\mathrm{H}}\left(\boldsymbol{I}+\boldsymbol{h}_2\boldsymbol{h}_2^{\mathrm{H}}q_2+\boldsymbol{h}_3\boldsymbol{h}_3^{\mathrm{H}}q_3\right)^{-1}\boldsymbol{h}_2\left(1+q_2\boldsymbol{h}_2^{\mathrm{H}}\left(\boldsymbol{I}+\boldsymbol{h}_3\boldsymbol{h}_3^{\mathrm{H}}q_3\right)^{-1}\boldsymbol{h}_2\right).\end{aligned} \tag{3.33}$$

Finally, substituting Eqs. 3.32 and 3.33 in Eq. 3.31 and after some trivial algebra we obtain

$$\begin{aligned}\log_2\left(1+\frac{\boldsymbol{h}_2^{\mathrm{H}}\boldsymbol{\Sigma}_2\boldsymbol{h}_2\boldsymbol{h}_2^{\mathrm{H}}\boldsymbol{\Sigma}_3\boldsymbol{h}_2}{\left(1+\boldsymbol{h}_2^{\mathrm{H}}\boldsymbol{\Sigma}_1\boldsymbol{h}_2+\boldsymbol{h}_2^{\mathrm{H}}\boldsymbol{\Sigma}_2\boldsymbol{h}_2\right)^2}\right) = \\ = \log_2\left(\frac{(1+\|\boldsymbol{h}_3\|_2^2q_3)(1+\|\boldsymbol{h}_2\|_2^2q_2)}{(1+\|\boldsymbol{h}_3\|_2^2q_3)(1+\|\boldsymbol{h}_2\|_2^2q_2)-q_2q_3|\boldsymbol{h}_2^{\mathrm{H}}\boldsymbol{h}_3|^2}\right).\end{aligned}$$

It remains to be shown that the left-hand side of Eq. 3.28 is equal to this expression. Using the identity $|\boldsymbol{I}+\boldsymbol{AB}| = |\boldsymbol{I}+\boldsymbol{BA}|$ we get

$$|\boldsymbol{I}_t + q_k\boldsymbol{h}_k\boldsymbol{h}_k^{\mathrm{H}}| = 1+\|\boldsymbol{h}_k\|_2^2q_k, \quad k=1,2. \tag{3.34}$$

Using this identity after successive application of the matrix inversion lemma to $\left(\boldsymbol{I}_t+q_2\boldsymbol{h}_2\boldsymbol{h}_2^{\mathrm{H}}+q_3\boldsymbol{h}_3\boldsymbol{h}_3^{\mathrm{H}}\right)^{-1}$ and $\left(\boldsymbol{I}_t+q_3\boldsymbol{h}_k\boldsymbol{h}_k^{\mathrm{H}}\right)^{-1}$ with $k=2$ or $k=3$ it is easy to show that

$$\left|\boldsymbol{I}_t+q_2\boldsymbol{h}_2\boldsymbol{h}_2^{\mathrm{H}}+q_3\boldsymbol{h}_3\boldsymbol{h}_3^{\mathrm{H}}\right|^{-1} = \left((1+\|\boldsymbol{h}_3\|_2^2q_3)(1+\|\boldsymbol{h}_2\|_2^2q_2)-q_2q_3|\boldsymbol{h}_2^{\mathrm{H}}\boldsymbol{h}_3|^2\right)^{-1}.$$

Substituting this equation and Eqs. 3.34 in the left-hand side of Eq. 3.28 the proof follows.

3.2.1 Memoryless channels

The first algorithm proposed in the literature to specifically solve Problem 3.23 was presented in [133]. This algorithm performs in each iteration a line search within the feasible region by moving along the direction of the principal eigenvalue of the gradient of the objective function. The main drawback of this algorithm is the high sensitivity of the convergence rate with respect to the dimensionality of the input space. In part motivated by this flaw, a number of algorithmic approaches to the weighted sum-rate maximization problem have appeared of late, which exhibit faster convergence. In [69] an algorithm is presented for the particular case of single-antenna receivers that builds upon the iterative waterfilling principle discussed in Section 3.1.1.1. A conjugate gradient projection algorithm for the general case is proposed in [75]. In [10] an algorithm has been presented that is based on a projected conjugate gradient approach and operates on the precoding filters. Finally, a gradient ascent projection algorithm is proposed in [62]. Here, we will first briefly review the initial algorithm proposed in [133], which has so far been the algorithm

of reference for weighted sum-rate maximization, and, then, we briefly expose the basic principles of the gradient ascent projection algorithm presented in [62], which shares some of its basics elements with the algorithms presented in [75] and [10].

Without loss of generality and for purposes of notational convenience, in the next two sections weights $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_K$ are assumed.

3.2.1.1 Rank-one gradient ascent

This algorithm, proposed in [133], computes in each iteration a new set of covariance matrices within the feasible region by moving along the direction of the principal eigenvalue of the gradient of the objective function. To be more specific, consider the objective function of Problem 3.23 under the assumption on the weights made above,

$$f(\boldsymbol{Q}_{1,\ldots,K}) = \sum_{k=1}^{K} \eta_k \log_2 \left(\left| \boldsymbol{I}_t + \sum_{j \geq k} \boldsymbol{H}_j^{\mathrm{H}} \boldsymbol{Q}_j \boldsymbol{H}_j \right| \right),$$

where $\eta_k = \mu_k - \mu_{k-1}$, $k = 2, \ldots, K$ and $\eta_1 = \mu_1$. The gradient of this function with respect to $\boldsymbol{Q}_i$ is given by

$$\boldsymbol{G}_i = \sum_{k=1}^{i} \eta_k \boldsymbol{H}_i \left(\boldsymbol{I}_t + \sum_{j \geq k} \boldsymbol{H}_j^{\mathrm{H}} \boldsymbol{Q}_j \boldsymbol{H}_j \right)^{-1} \boldsymbol{H}_i^{\mathrm{H}}. \tag{3.35}$$

Now, consider the vector resulting from

$$\boldsymbol{u} = \arg\max_{\boldsymbol{v}} \left\{ \boldsymbol{v}^{\mathrm{H}} \boldsymbol{G}_k \boldsymbol{v} | k = 1, \ldots, K \right\}, \tag{3.36}$$

subject to $\|\boldsymbol{v}\| = 1$. This vector is the eigenvector associated with the largest principal eigenvalue of all gradient matrices. Let $\boldsymbol{Q}_{1,\ldots,K}^{(\ell)}$ be the set of covariance matrices obtained at the ℓth iteration of the algorithm and $\boldsymbol{u}^{(\ell)}$ the maximizer of Problem 3.36 for these matrices. Assume that $\boldsymbol{u}^{(\ell)}$ is the principal eigenvector of $\boldsymbol{G}_i$, i.e., the maximum in Problem 3.36 is reached for index $k = i$. The update rule for the covariance matrices is given by

$$\begin{aligned} \boldsymbol{Q}_k^{(\ell+1)} &= (1-\alpha)\boldsymbol{Q}_k^{(\ell)} + P\alpha \boldsymbol{u}^{(\ell)} \boldsymbol{u}^{(\ell),\mathrm{H}}, && k = i, \\ \boldsymbol{Q}_k^{(\ell+1)} &= (1-\alpha)\boldsymbol{Q}_k^{(\ell)}, && k \neq i, \end{aligned}$$

where $\alpha \in (0,1)$ is the step size that can be optimized for each update. On the one hand, this update preserves the positive semidefinite property of the covariance matrices. On the other hand, it can be easily verified that if matrices $\boldsymbol{Q}_{1,\ldots,K}^{(\ell)}$ satisfy the transmit power constraint, $\boldsymbol{Q}_{1,\ldots,K}^{(\ell+1)}$ also satisfy this constraint. That there is always a step size for which this update yields an increase of the objective function can be shown by analyzing the linear approximation of $f(\boldsymbol{Q}_{1,\ldots,K}^{(\ell+1)})$ around $\boldsymbol{Q}_{1,\ldots,K}^{(\ell)}$, i.e.,

$$f(\boldsymbol{Q}_{1,\ldots,K}^{(\ell+1)}) \approx f(\boldsymbol{Q}_{1,\ldots,K}^{(\ell)}) - \alpha \sum_{k=1}^{K} \mathrm{Tr}\left\{ \boldsymbol{G}_k \boldsymbol{Q}_k^{(\ell)} \right\} + P\alpha \mathrm{Tr}\left\{ \boldsymbol{G}_i \boldsymbol{u}^{(\ell)} \boldsymbol{u}^{(\ell),\mathrm{H}} \right\}. \tag{3.37}$$

The third term on the right-hand side is equal to $\alpha P\lambda$, where λ is the eigenvalue associated to $\boldsymbol{u}^{(\ell)}$. The second term is upper bounded by $\alpha P\lambda$. This can be easily shown by noting

$$\sum_{k=1}^{K} \mathrm{Tr}\left\{\boldsymbol{G}_k \boldsymbol{Q}_k^{(\ell)}\right\} = \sum_{k=1}^{K} \mathrm{Tr}\left\{\boldsymbol{\Lambda}_k \hat{\boldsymbol{Q}}_k^{(\ell)}\right\} = \sum_{k=1}^{K}\sum_{j=1}^{J} [\boldsymbol{\Lambda}_k]_{j,j} \left[\hat{\boldsymbol{Q}}_k^{(\ell)}\right]_{j,j} \leq \lambda \sum_{k=1}^{K} \mathrm{Tr}\left\{\hat{\boldsymbol{Q}}_k^{(\ell)}\right\} \leq \lambda P,$$

where $\boldsymbol{\Lambda}_k$ is the matrix of eigenvalues of $\boldsymbol{G}_k$, $\boldsymbol{G}_k = \boldsymbol{U}_k \boldsymbol{\Lambda}_k \boldsymbol{U}_k^{\mathrm{H}}$ and $\hat{\boldsymbol{Q}}_k^{(\ell)} = \boldsymbol{U}_k^{\mathrm{H}} \boldsymbol{Q}_k^{(\ell)} \boldsymbol{U}_k$. Thus, the right-hand side of Eq. 3.37 is never smaller than $f(\boldsymbol{Q}_{1,\ldots,K}^{(\ell)})$. Choosing α small enough the linear approximation can be done arbitrarily accurate and, hence, at least for some of $\alpha > 0$, $f(\boldsymbol{Q}_{1,\ldots,K}^{(\ell+1)})$ will necessarily be greater than $f(\boldsymbol{Q}_{1,\ldots,K}^{(\ell)})$. The pseudocode corresponding to this approach is given in Algorithm 3.8. A major flaw of this algorithm is that only the structure of the covariance matrix of one user is updated in each iteration. As a result, the convergence speed of the algorithm strongly depends on the number of users in the system or, more precisely, on the number of users that are eventually served (cf. [69]).

Algorithm 3.8 Rank-one gradient ascent

$\boldsymbol{Q}_k^{(0)} \leftarrow \frac{P}{Kr_k}\boldsymbol{I}_{r_k},\ k = 1,\ldots,K,\quad \ell \leftarrow 0$
$R^{\mathrm{new}} \leftarrow 0$
repeat
  $R^{\mathrm{old}} \leftarrow R^{\mathrm{new}}$
  for $m = 1$ to K **do**
    $\boldsymbol{G}_i^{(\ell)} \leftarrow \sum_{k=1}^{i} \eta_k \boldsymbol{H}_i \left(\boldsymbol{I}_t + \sum_{j\geq k} \boldsymbol{H}_j^{\mathrm{H}} \boldsymbol{Q}_j^{(\ell)} \boldsymbol{H}_j\right)^{-1} \boldsymbol{H}_i^{\mathrm{H}},\quad i = 1,\ldots,K$
  end for
  $\{\boldsymbol{u}^{(\ell)}, j\} \leftarrow \underset{\boldsymbol{v},k=1,\ldots,K}{\arg\max}\ \boldsymbol{v}^{\mathrm{H}} \boldsymbol{G}_k^{(\ell)} \boldsymbol{v},$ subject to $\|v\| = 1.$
  $\boldsymbol{Q}_k^{(\ell+1)} = (1-\alpha)\boldsymbol{Q}_k^{(\ell)} + \delta[k-j]P\alpha \boldsymbol{u}^{(\ell)} \boldsymbol{u}^{(\ell),\mathrm{H}},\quad k = 1,\ldots,K$
  $\ell \leftarrow \ell + 1$
  $R^{\mathrm{new}} \leftarrow f(\boldsymbol{Q}_{1,\ldots,K}^{(\ell)})$
until $R^{\mathrm{new}} - R^{\mathrm{old}} < \epsilon$

3.2.1.2 Projected gradient ascent

In each iteration, this algorithm, which has been proposed in [62], moves along the direction indicated by the gradient of the objective function in order to search for an update of the covariance matrices. In contrast to the rank-one gradient ascent algorithm discussed in the previous section, here, the feasible region might be abandoned during the update operation. Therefore, a projection step is needed that maps the updated set of matrices onto a set of new feasible covariance matrices. The principle is rather general. However, the implementation specifics for this particular application exhibit some interesting features.

Assume that, after the ℓth iteration, covariance matrices $\boldsymbol{Q}_{1,\ldots,K}^{(\ell)}$ are obtained. In order to compute an improved set of covariance matrices, first, the gradients of the objective

function are computed as indicated in Eq. 3.35. Then, an update of the covariance matrices is done according to the following rule

$$\hat{\boldsymbol{Q}}_k^{(\ell+1)} = \boldsymbol{Q}_k^{(\ell)} + \alpha \boldsymbol{G}_k, \quad k = 1, \ldots, K.$$

where $\alpha > 0$ is the step size. Contrary to the update rule of Algorithm 3.8, this update rule modifies the structure of all covariance matrices in the direction of the gradient. This is basically the reason for the superior convergence performance of this approach. However, now, the total power constraint might be violated after such an update.[2] In order to obtain a feasible set of covariance matrices out of $\hat{\boldsymbol{Q}}_{1,\ldots,K}^{(\ell+1)}$ a projection must be performed onto the feasibility region. Convergence is guaranteed if an orthogonal projection is chosen, i.e., if for a given set of matrices $\hat{\boldsymbol{Q}}_{1,\ldots,K}^{(\ell+1)}$, the set of feasible matrices is computed that lying within the feasible region are closest to the matrices $\hat{\boldsymbol{Q}}_{1,\ldots,K}^{(\ell+1)}$ according to a certain norm. Using the Frobenius norm the set of covariance matrices that fulfil the power constraint and are closest to $\hat{\boldsymbol{Q}}_{1,\ldots,K}^{(\ell+1)}$ can be computed as [75, 62]

$$\boldsymbol{Q}_k^{(\ell+1)} = [\hat{\boldsymbol{Q}}_k^{(\ell+1)} - \xi \boldsymbol{I}_{r_k}]^+, \quad k = 1, \ldots, K$$

where $\xi \geq 0$ is chosen so that the power constraint is fulfilled and $[\bullet]^+$ is an operator that sets negative eigenvalues equal to zero. In the same way as the rate maximizing power distribution over parallel channels can be visualized as waterfilling a recipient with different levels, this solution admits an waterspilling interpretation, according to which water is let out from a recipient as long as the total volume of water contained in the recipient exceeds a desired value [62]. A sketch of this procedure is given in Algorithm 3.9.

Algorithm 3.9 Projected gradient ascent

1: $\boldsymbol{Q}_k^{(0)} \leftarrow \frac{P}{K r_k} \boldsymbol{I}_{r_k},\ k = 1, \ldots, K, \quad \ell \leftarrow 0$
2: $R^{\text{new}} \leftarrow 0$
3: **repeat**
4: $\quad R^{\text{old}} \leftarrow R^{\text{new}}$
5: $\quad$ **for** $m = 1$ to K **do**
6: $\quad\quad \boldsymbol{G}_i^{(\ell)} \leftarrow \sum_{k=1}^{i} \eta_k \boldsymbol{H}_i \left(\boldsymbol{I}_t + \sum_{j \geq k} \boldsymbol{H}_j^{\text{H}} \boldsymbol{Q}_j^{(\ell)} \boldsymbol{H}_j \right)^{-1} \boldsymbol{H}_i^{\text{H}}, \quad i = 1, \ldots, K$
7: $\quad$ **end for**
8: $\quad \hat{\boldsymbol{Q}}_k^{(\ell+1)} \leftarrow \boldsymbol{Q}_k^{(\ell)} + \alpha \boldsymbol{G}_k^{(\ell)}, \quad k = 1, \ldots, K$
9: $\quad \boldsymbol{Q}_k^{(\ell+1)} \leftarrow [\hat{\boldsymbol{Q}}_k^{(\ell+1)} - \xi \boldsymbol{I}_{r_k}]^+, \quad k = 1, \ldots, K$
10: $\quad \ell \leftarrow \ell + 1$
11: $\quad R^{\text{new}} \leftarrow f(\boldsymbol{Q}_{1,\ldots,K}^{(\ell)})$
12: **until** $R^{\text{new}} - R^{\text{old}} < \epsilon$

[2] Note that due to the fact that $\boldsymbol{G}_k \geq 0$, the positive semidefinite constraint can never be violated by this update rule.

3.2.2 Time-dispersive channels

The algorithms reviewed in the last section can be straightforwardly applied to the MIMO OFDM multiuser model given by Eqs. 3.13. As already discussed in Section 3.1.2, without consideration of the block diagonal structure of the channels, this trivial extension has the shortcoming of a cubic complexity order in the number of subcarriers. Nevertheless, along the lines of the discussion in Section 3.1.2, it can be shown that also for the weighted sum-rate optimization problem a block diagonal structure of the covariance matrix of each user is optimum [116]. As a result, for the MIMO OFDM broadcast channel, Problem 3.23 can be restated as

$$\max_{\substack{\boldsymbol{Q}_{1,\ldots,K}\\1,\ldots,N}} \sum_{n=1}^{N}\sum_{k=1}^{K} \eta_k \log_2 \left(\left| \boldsymbol{I}_t + \sum_{j=k}^{K} \boldsymbol{H}_{j,n}^{\mathrm{H}} \boldsymbol{Q}_{j,n} \boldsymbol{H}_{j,n} \right| \right), \tag{3.38}$$

subject to $\sum_{k=1,n=1}^{K,N} \mathrm{Tr}\{\boldsymbol{Q}_{k,n}\} \leq P$ and $\boldsymbol{Q}_{k,n} \geq \boldsymbol{0},\ \forall k,n$. Here, $\boldsymbol{Q}_{k,n}$ denotes the covariance matrix of user k on subcarrier n. In the next section a decomposition approach is presented that makes possible the application of any weighted sum-rate maximizing approach for memoryless channels to Problem 3.38 with linear complexity in the number of subcarriers. This procedure is based on the factorization of the covariance matrices as the product of a normalized covariance matrix and a scalar representing the power used on the corresponding subcarrier.

3.2.2.1 Factorization-based decomposition approach

This algorithm was introduced in [115, 116] in order to circumvent the dramatic slowdown in convergence speed experienced by Algorithm 3.8 for increasing number of subcarriers when directly applied to solve Problem 3.38. The main merit of the algorithm is that if applied in combination with Algorithm 3.8 the number of required iterations to reach convergence becomes independent of the number of subcarriers. The recently appeared approaches in [75, 10, 62] admit simple extensions that solve Problem 3.38 in linear time in the number of subcarriers and have a convergence behavior that is essentially insensitive to the number of subcarriers in the system. This somehow undermines the significance of the method described in the sequel. Nonetheless, this approach has a universal character regarding the kind of algorithms it can be combined with. Furthermore, the general way of proceeding might be of interest in contexts other than weighted sum-rate maximization.

For each subcarrier, we factorize $\boldsymbol{Q}_{k,n} = P_n \hat{\boldsymbol{Q}}_{k,n}$ such that $\sum_{k=1}^{K} \mathrm{Tr}\{\hat{\boldsymbol{Q}}_{k,n}\} \leq 1$ and $\sum_{n=1}^{N} P_n \leq P$. Taking this factorization into account, optimum covariance matrices are found iterating the following two steps. First, for given $\boldsymbol{p} = [\ P_1 \ \cdots \ P_N\ ]^{\mathrm{T}}$, solve

$$\max_{\hat{\boldsymbol{Q}}_{1,\ldots,K,n}} \sum_{k=1}^{K} \eta_k \log_2 \left(\left| \boldsymbol{I}_t + P_n \sum_{j=k}^{K} \boldsymbol{H}_{j,n}^{\mathrm{H}} \hat{\boldsymbol{Q}}_{j,n} \boldsymbol{H}_{j,n} \right| \right),$$

subject to $\sum_{k=1}^{K} \mathrm{Tr}\{\hat{\boldsymbol{Q}}_{k,n}\} \leq 1$ and $\hat{\boldsymbol{Q}}_{k,n} \geq \boldsymbol{0},\ \forall k$, for every n. Second, for a given set of

covariance matrices $\hat{\boldsymbol{Q}}_{k,n}$, $k = 1, \ldots, K$, $n = 1, \ldots, N$, solve

$$\max_{\boldsymbol{p}} \sum_{n=1}^{N} \sum_{k=1}^{K} \eta_k \log_2 \left(\left| \boldsymbol{I}_t + P_n \sum_{j=k}^{K} \boldsymbol{H}_{j,n}^{\mathrm{H}} \hat{\boldsymbol{Q}}_{j,n} \boldsymbol{H}_{j,n} \right| \right), \tag{3.39}$$

subject to $\sum_{n=1}^{N} P_n \leq P$ and $P_n \geq 0$. Both problems are convex. In the second step, an optimum power allocation over subcarriers $\boldsymbol{p}$ is found for a given set of normalized covariance matrices. In the first, given the optimum power allocation $\boldsymbol{p}$ obtained in the previous iteration, an optimum set of normalized covariance matrices is found for every subcarrier. It is clear that each step improves the value of the objective function in Eq. 3.38 and, hence, convergence is guaranteed.

In the first step, optimization of normalized covariance matrices can be done applying any of the existing weighted sum-rate maximizing algorithms for memoryless channels. In the second step, the KKT conditions of Problem 3.39 yield the following set of equations,

$$\sum_{k=1}^{K} \eta_k \mathrm{Tr} \left\{ \left(\boldsymbol{I}_t + P_n \boldsymbol{A}_{k,n} \right)^{-1} \boldsymbol{A}_{k,n} \right\} - \nu + \xi_n = 0, \ \forall n, \tag{3.40}$$

$$P - \sum_{n=1}^{N} P_n \geq 0, \ \nu \geq 0, \ P_n \geq 0, \ \xi_n \geq 0, \ \forall n,$$

$$\nu \left(P - \sum_{n=1}^{N} P_n \right) = 0, \ \xi_n P_n = 0, \ \forall n,$$

where $\boldsymbol{A}_{k,n} = \sum_{j=k}^{K} \boldsymbol{H}_{j,n}^{\mathrm{H}} \hat{\boldsymbol{Q}}_{j,n} \boldsymbol{H}_{j,n}$. Considering the eigenvalues $\lambda_{k,n}^{s}$, $s = 1, \ldots, t$, of matrix $\boldsymbol{A}_{k,n}$, Eq. 3.40 can be rewritten as

$$\sum_{k=1}^{K} \sum_{s=1}^{t} \frac{\eta_k \lambda_{k,n}^{s}}{1 + P_n \lambda_{k,n}^{s}} - \nu + \xi_n = 0, \ \forall n.$$

An efficient algorithm can be implemented that computes the power allocation $\boldsymbol{p}$ satisfying these conditions based on the following two observations.

Observation 1. *For a given ν, $P_n \neq 0$ if and only if $\sum_{k=1}^{K} \sum_{s=1}^{t} \eta_k \lambda_{k,n}^{s} > \nu$. In that case, $\xi_n = 0$ and*

$$\sum_{k=1}^{K} \sum_{s=1}^{t} \frac{\eta_k \lambda_{k,n}^{s}}{1 + P_n \lambda_{k,n}^{s}} - \nu$$

is a monotonically decreasing function of the transmit power P_n.

Observation 2. *The optimum ν is a monotonically decreasing function of the transmit power P. In addition,*

$$\nu < \max_{n} \left\{ \sum_{k=1}^{K} \sum_{s=1}^{t} \eta_k \lambda_{k,n}^{s} \right\},$$

i.e., at least one subcarrier gets some power.

From Observation 1 it becomes clear that for a given ν there is a unique power allocation $\boldsymbol{p}$ which can be efficiently computed. On the other hand, according to Observation 2, if this power allocation exceeds the available transmit power, ν should be increased, otherwise it should be decreased. In this way, bisection can be used in order to compute ν corresponding to the particular transmit power constraint. The pseudocode for this decomposition approach is given in Algorithm 3.10. The complexity order per iteration is linear in the number of subcarriers as the weighted sum-rate maximization problem in line 7 must be solved once per subcarrier. This can be done by applying any of the existing approaches for weighted sum-rate maximization in memoryless channels. Note that the complexity order involved in the computation of the optimum power allocation in line 9 is also linear in the number of subcarriers.

Algorithm 3.10 Factorization-based decomposition approach

1: $\hat{\boldsymbol{Q}}_{k,n}^{(0)} \leftarrow \frac{1}{Kr_k}\boldsymbol{I}_{r_k},\ k=1,\ldots,K,\ n=1,\ldots,N$
2: $P_n^{(0)} = P/N,\ n=1,\ldots,N,\quad \ell \leftarrow 0$
3: $R^{\text{new}} \leftarrow 0$
4: **repeat**
5: $\quad R^{\text{old}} \leftarrow R^{\text{new}}$
6: $\quad$ **for** $n=1$ to N **do**
7: $\quad\quad \hat{\boldsymbol{Q}}_{1,\ldots,K,n}^{(\ell+1)} \leftarrow \underset{\hat{\boldsymbol{Q}}_{1,\ldots,K,n}}{\arg\max} \sum_{k=1}^{K} \eta_k \log\left(\left|\boldsymbol{I}_t + P_n^{(\ell)} \sum_{j=k}^{K} \boldsymbol{H}_{j,n}^{\mathrm{H}} \hat{\boldsymbol{Q}}_{j,n} \boldsymbol{H}_{j,n}\right|\right)$
$\quad\quad$ subject to $\sum_{k=1}^{K} \mathrm{Tr}\{\bar{\boldsymbol{Q}}_{k,n}\} \leq 1,\ \hat{\boldsymbol{Q}}_{k,n} \geq \boldsymbol{0},\ \forall k$
8: $\quad$ **end for**
9: $\quad \boldsymbol{p}^{(\ell+1)} \leftarrow \underset{\boldsymbol{p}}{\arg\max} \sum_{n=1}^{N}\sum_{k=1}^{K} \eta_k \log\left(\left|\boldsymbol{I}_t + P_n \sum_{j=k}^{K} \boldsymbol{H}_{j,n}^{\mathrm{H}} \hat{\boldsymbol{Q}}_{j,n}^{(\ell+1)} \boldsymbol{H}_{j,n}\right|\right)$
$\quad$ subject to $\sum_{n=1}^{N} P_n \leq P,\ P_n \geq 0$
10: $\quad \ell \leftarrow \ell + 1$
11: $\quad R^{\text{new}} \leftarrow \sum_{n=1}^{N}\sum_{k=1}^{K} \eta_k \log\left(\left|\boldsymbol{I}_t + P_n^{(\ell)} \sum_{j=k}^{K} \boldsymbol{H}_{j,n}^{\mathrm{H}} \hat{\boldsymbol{Q}}_{j,n}^{(\ell)} \boldsymbol{H}_{j,n}\right|\right)$
12: **until** $R^{\text{new}} - R^{\text{old}} < \epsilon$

3.3 Rate balancing

Weighted sum-rate maximization is a suitable policy in the context of communication systems with stationary random arrival of information and buffering capability [8, 9]. In such systems, if, at each time slot, the priorities are chosen to be proportional to the length of the queue corresponding to each user, the system can be stabilized, i.e., the average delay is bounded for all users. However, in the case of very stringent delay constraints and limited mobility, assigning priorities to users and optimizing weighted sum rate does not guarantee that the final ranking of the users, as given by the rates they obtain, corresponds to the intended prioritization. For instance, it may happen that a high priority user obtains far a

lower rate than a low priority user. This is generally the case if the channel of the latter is good enough as compared to the channel of the former. Sometimes, it might be desirable to have a stronger control upon the relative performance achieved by the users in the network with respect to each other. Here is where the rate balancing problem formulation becomes relevant. Now, the users are assigned relative rates q_k, $k = 1, \ldots, K$, rather than priorities. A relative rate expresses the share that each user should get out of the total transmitted rate. Mathematically, the optimization problem can be written as

$$\max_{\gamma,\boldsymbol{\rho}} \gamma \quad \text{subject to} \quad \gamma \boldsymbol{q} \leq \boldsymbol{\rho}, \ \forall k, \qquad \boldsymbol{\rho} \in \mathcal{R}^{\mathrm{DPC}}(P), \tag{3.41}$$

where $\mathcal{R}^{\mathrm{DPC}}(P)$ is the BC capacity region for a particular channel realization as defined in Eq. 2.23 or Eq. 2.24, $\boldsymbol{\rho} = [\ R_1 \ \cdots \ R_K\ ]^{\mathrm{T}}$ and $\boldsymbol{q} = [\ q_1 \ \cdots \ q_K\ ]^{\mathrm{T}}$. Obviously, due to duality, $\mathcal{R}^{\mathrm{DPC}}(P)$ may be replaced by $\mathcal{R}^{\mathrm{MAC}}(P)$ in the problem statement. Due to convexity of the capacity region, the maximum of Problem 3.41 can be achieved with equality in the constraints. That is, this problem is equivalent to finding the intersection between the straight line defined by the constraint $\gamma \boldsymbol{q} = \boldsymbol{\rho}$ and the boundary of the capacity region.

As the capacity region is a convex set, Problem 3.41 is also convex. Furthermore, the feasibility region has always a non-empty interior and, therefore, strong duality holds [13]. Consequently, a solution to the rate-balancing problem can be found by solving the dual minimization problem. The Lagrangian dual function of Problem 3.41 with respect to the constraint $\gamma \boldsymbol{q} \leq \boldsymbol{\rho}$ can be written as

$$g(\boldsymbol{\mu}) = \sup_{\gamma,\boldsymbol{\rho}} \gamma + \sum_{k=1}^{K} \mu_k \left(\frac{R_k}{q_k} - \gamma \right), \tag{3.42}$$

subject to $\mu_k \geq 0$, $\forall k$, and $\boldsymbol{\rho} \in \mathcal{R}^{\mathrm{DPC}}(P)$, where $\boldsymbol{\mu} = [\ \mu_1 \ \cdots \ \mu_K\ ]^{\mathrm{T}}$. This function is equal to ∞ unless $\mu_1 + \cdots + \mu_K = 1$. As a result, the dual problem can be written as

$$\min_{\boldsymbol{\mu}} \max_{\boldsymbol{\rho}} \sum_{k=1}^{K} \mu_k \frac{R_k}{q_k},$$

subject to $\|\boldsymbol{\mu}\|_1 = 1$, $\mu_k \geq 0$, $\forall k$ and $\boldsymbol{\rho} \in \mathcal{R}^{\mathrm{DPC}}(P)$. Alternatively, the first constraint can be incorporated into the objective function in order to obtain [71]

$$\min_{\tilde{\boldsymbol{\mu}}} \max_{\boldsymbol{\rho}} \frac{R_K}{q_K} + \sum_{k=1}^{K-1} \mu_k \left(\frac{R_k}{q_k} - \frac{R_K}{q_K} \right), \tag{3.43}$$

subject to $\|\tilde{\boldsymbol{\mu}}\|_1 \leq 1$, $\mu_k \geq 0$, $\forall k$ and $\boldsymbol{\rho} \in \mathcal{R}^{\mathrm{DPC}}(P)$, where $\tilde{\boldsymbol{\mu}} = [\ \mu_1 \ \cdots \ \mu_{K-1}\ ]^{\mathrm{T}}$. Let $\bar{\boldsymbol{\rho}} = [\ \bar{R}_1 \ \cdots \ \bar{R}_K\ ]^{\mathrm{T}}$ be a maximizer of

$$g(\tilde{\boldsymbol{\mu}}) = \max_{\boldsymbol{\rho}} \frac{R_K}{q_K} + \sum_{k=1}^{K-1} \mu_k \left(\frac{R_k}{q_k} - \frac{R_K}{q_K} \right), \tag{3.44}$$

subject to $\boldsymbol{\rho} \in \mathcal{R}^{\mathrm{DPC}}(P)$, for given $\tilde{\boldsymbol{\mu}}$. From this definition of $g(\tilde{\boldsymbol{\mu}})$ it immediately follows

$$g(\tilde{\boldsymbol{\mu}} + \Delta\tilde{\boldsymbol{\mu}}) - g(\tilde{\boldsymbol{\mu}}) \geq \sum_{k=1}^{K-1} \left(\frac{\bar{R}_k}{q_k} - \frac{\bar{R}_K}{q_K} \right) \Delta\mu_k,$$

i.e., the vector $\boldsymbol{s} = [\ S_1 \ \cdots \ S_{K-1}\ ]^{\mathrm{T}}$ with

$$S_k = \frac{\bar{R}_k}{q_k} - \frac{\bar{R}_K}{q_K}$$

is a subgradient of $g(\tilde{\boldsymbol{\mu}})$ at the given $\tilde{\boldsymbol{\mu}}$. Different from some of the algorithms based on Lagrangian duality discussed so far, where bisection could be applied to find the optimum in the one-dimensional dual input space, now, the dual variable $\tilde{\boldsymbol{\mu}}$ is, in general, multidimensional. This calls for some more sophisticated subgradient-based methods. In the next sections we briefly discuss two of them. The ellipsoid method, applied to this problem in [71], and a projected subgradient method.

3.3.1 Ellipsoid method

The ellipsoid method was introduced in the late seventies. Its theoretical relevance was soon revealed as, shortly after its introduction, it made possible to show that linear programs are solvable in polynomial time, which had been a long-standing open question until then (cf. [86] and references therein). This algorithm can be classified as a localization method, to which cutting-plane methods also belong [12]. Starting from an initial point, at each step these methods make use of a gradient or subgradient in order to discard part of the search space from consideration in all future steps. In this way the search space reduces at each step until the optimum point is localized within a set whose dimensions do not exceed the desired accuracy. Compared to other localization methods, the ellipsoid method shows a slow convergence but iterations are very simple. This algorithm was applied to the rate balancing problem in [71].

An ellipsoid in $\mathbb{R}^{K-1}$ can be written as

$$\mathcal{E}(\tilde{\boldsymbol{\mu}}^{(0)}, \boldsymbol{E}^{(0)}) = \left\{ \boldsymbol{z} | (\boldsymbol{z} - \tilde{\boldsymbol{\mu}}^{(0)})^{\mathrm{T}} (\boldsymbol{E}^{(0)})^{-1} (\boldsymbol{z} - \tilde{\boldsymbol{\mu}}^{(0)}) \leq 1 \right\},$$

where $\boldsymbol{z} \in \mathbb{R}^{K-1}$ and $\boldsymbol{E}^{(0)}$ is a symmetric $K-1 \times K-1$ matrix with real entries whose eigenvalues represent the square of the lengths of the semi-axes of the ellipsoid. The center of the ellipsoid is given by $\tilde{\boldsymbol{\mu}}^{(0)}$. In the first step, an ellipsoid must be computed that comprises all the feasible region, e.g., $\boldsymbol{E}^{(0)} = (1 - 1/K)\boldsymbol{I}_{K-1}$, $\mu_1^{(0)} = \cdots = \mu_{K-1}^{(0)} = 1/K$. Obviously, the search can be restricted to points within this first ellipsoid as the optimum point is certain to be a feasible point. Assume that after ℓ iterations, we know that the optimum is in an ellipsoid $\mathcal{E}(\tilde{\boldsymbol{\mu}}^{(\ell)}, \boldsymbol{E}^{(\ell)})$. If the point $\tilde{\boldsymbol{\mu}}^{(\ell)}$ is feasible and we let $\boldsymbol{s}^{(\ell)}$ be the subgradient of $g(\tilde{\boldsymbol{\mu}})$ at $\tilde{\boldsymbol{\mu}}^{(\ell)}$, using the definition of subgradient, it can be easily shown that the optimum necessarily lies in the set

$$\mathcal{E}(\tilde{\boldsymbol{\mu}}^{(\ell)}, \boldsymbol{E}^{(\ell)}) \cap \{\boldsymbol{z} | \boldsymbol{s}^{(\ell),\mathrm{T}} (\boldsymbol{z} - \tilde{\boldsymbol{\mu}}^{(\ell)}) \leq 0\}.$$

In the next iteration the algorithm computes the ellipsoid $\mathcal{E}(\tilde{\boldsymbol{\mu}}^{(\ell+1)}, \boldsymbol{E}^{(\ell+1)})$ of minimum volume enclosing this set as [12]

$$\tilde{\boldsymbol{\mu}}^{(\ell+1)} = \tilde{\boldsymbol{\mu}}^{(\ell)} - \frac{1}{K}\boldsymbol{E}^{(\ell)}\tilde{\boldsymbol{s}}^{(\ell)}, \tag{3.45}$$

$$\boldsymbol{E}^{(\ell+1)} = \frac{(K-1)^2}{(K-1)^2-1}\left(\boldsymbol{E}^{(\ell)} - \frac{2}{K}\boldsymbol{E}^{(\ell)}\tilde{\boldsymbol{s}}^{(\ell)}\tilde{\boldsymbol{s}}^{(\ell),\mathrm{T}}\boldsymbol{E}^{(\ell)}\right), \tag{3.46}$$

where $\tilde{\boldsymbol{s}}^{(\ell)} = \boldsymbol{s}^{(\ell)}/\sqrt{\boldsymbol{s}^{(\ell),\mathrm{T}}\boldsymbol{E}^{(\ell)}\boldsymbol{s}^{(\ell)}}$. A case that was somehow neglected in [71] is the occurrence of an infeasible $\tilde{\boldsymbol{\mu}}^{(\ell)}$ at the end of an iteration. Let $\boldsymbol{s}_k^{(\ell)}$, $k = 1, \ldots, K-1$, be the gradients[3] of the constraints $f_k(\tilde{\boldsymbol{\mu}}) = -\mu_k \leq 0$, $k = 1, \ldots, K-1$, and $f_K(\tilde{\boldsymbol{\mu}}) = \mu_1 + \cdots + \mu_{K-1} \leq 1$, and assume that constraint $f_k(\tilde{\boldsymbol{\mu}}) \leq 0$ is violated by $\tilde{\boldsymbol{\mu}}^{(\ell)}$, i.e., $f_k(\tilde{\boldsymbol{\mu}}^{(\ell)}) > 0$. From the definition of subgradient, it can be shown that, under these assumptions, the optimum necessarily lies in

$$\mathcal{E}(\tilde{\boldsymbol{\mu}}^{(\ell)}, \boldsymbol{E}^{(\ell)}) \cap \{\boldsymbol{z} | f_k(\tilde{\boldsymbol{\mu}}^{(\ell)}) + \boldsymbol{s}_k^{(\ell),\mathrm{T}}(\boldsymbol{z} - \tilde{\boldsymbol{\mu}}^{(\ell)}) \leq 0\}.$$

In this case the center and matrix of the ellipsoid of minimum volume enclosing this set is given by [12]

$$\tilde{\boldsymbol{\mu}}^{(\ell+1)} = \tilde{\boldsymbol{\mu}}^{(\ell)} - \frac{1+\alpha(K-1)}{K}\boldsymbol{E}^{(\ell)}\tilde{\boldsymbol{s}}_k^{(\ell)}, \tag{3.47}$$

$$\boldsymbol{E}^{(\ell+1)} = \frac{(K-1)^2(1-\alpha^2)}{(K-1)^2-1}\left(\boldsymbol{E}^{(\ell)} - \frac{2(1+\alpha(K-1))}{(1+\alpha)K}\boldsymbol{E}^{(\ell)}\tilde{\boldsymbol{s}}_k^{(\ell)}\tilde{\boldsymbol{s}}_k^{(\ell),\mathrm{T}}\boldsymbol{E}^{(\ell)}\right), \tag{3.48}$$

where $\tilde{\boldsymbol{s}}_k^{(\ell)} = \boldsymbol{s}_k^{(\ell)}/\sqrt{\boldsymbol{s}_k^{(\ell),\mathrm{T}}\boldsymbol{E}^{(\ell)}\boldsymbol{s}_k^{(\ell)}}$ and $\alpha = f_k(\tilde{\boldsymbol{\mu}}^{(\ell)})/\sqrt{\boldsymbol{s}_k^{(\ell),\mathrm{T}}\boldsymbol{E}^{(\ell)}\boldsymbol{s}_k^{(\ell)}}$. A sketch of the pseudocode corresponding to this method is given in Algorithm 3.11.

3.3.2 Projected subgradient method

This method is a generalization of the gradient descent projection method to convex nondifferentiable functions. Given a subgradient $\boldsymbol{s}^{(\ell)}$ of $g(\tilde{\boldsymbol{\mu}})$ at $\tilde{\boldsymbol{\mu}}^{(\ell)}$, the update rule is given by

$$\hat{\tilde{\boldsymbol{\mu}}}^{(\ell+1)} = \tilde{\boldsymbol{\mu}}^{(\ell)} - \alpha_\ell \boldsymbol{s}^{(\ell)},$$

where α_ℓ is the step size at the ℓth iteration. Of course, this update might result in a new $\hat{\tilde{\boldsymbol{\mu}}}^{(\ell+1)}$ that is not feasible. In this case, in order to restore feasibility this new point must be projected back onto the feasibility region. The minimum Euclidean distance projection of $\hat{\tilde{\boldsymbol{\mu}}}^{(\ell+1)}$ onto the feasibility region can be computed by solving

$$\min \|\boldsymbol{d}\|_2^2, \quad \text{subject to} \quad \begin{array}{l} \|\hat{\tilde{\boldsymbol{\mu}}}^{(\ell+1)} + \boldsymbol{d}\|_1 \leq 1, \\ \hat{\mu}_k + d_k \geq 0, \ k = 1, \ldots, K-1 \end{array} \tag{3.49}$$

[3] Due to the fact that the constraint functions are differentiable the concepts of subgradient and gradient are equivalent.

Algorithm 3.11 Ellipsoid method

$\mu_k^{(0)} \leftarrow \frac{1}{K},\ k = 1, \ldots, K-1$, $\boldsymbol{E}^{(0)} \leftarrow (1 - 1/K)\boldsymbol{I}_{K-1}$, $\ell \leftarrow 0$
Compute $g(\tilde{\boldsymbol{\mu}}^{(0)})$
repeat
  if $f_k(\tilde{\boldsymbol{\mu}}^{(\ell)}) \leq 0, \forall k$ **then**
    Compute $\tilde{\boldsymbol{\mu}}^{(\ell+1)}$ as in Eq. 3.45
    Compute $\boldsymbol{E}^{(\ell+1)}$ as in Eq. 3.46
  else
    Select violated constraint
    Compute $\tilde{\boldsymbol{\mu}}^{(\ell+1)}$ as in Eq. 3.47
    Compute $\boldsymbol{E}^{(\ell+1)}$ as in Eq. 3.48
  end if
  $\ell \leftarrow \ell + 1$
  Compute $g(\tilde{\boldsymbol{\mu}}^{(\ell)})$
until $|g(\tilde{\boldsymbol{\mu}}^{(\ell)}) - g(\tilde{\boldsymbol{\mu}}^{(\ell-1)})| < \epsilon$

The projected update is then given by $\tilde{\boldsymbol{\mu}}^{(\ell+1)} = \hat{\tilde{\boldsymbol{\mu}}}^{(\ell+1)} + \boldsymbol{d}$. Solving the KKT conditions for Problem 3.49 the following projection rule can be derived

$$\mu_k^{(\ell+1)} = \left[\left[\hat{\mu}_k^{(\ell+1)}\right]^+ - \eta\right]^+, \quad k = 1, \ldots, K-1$$

where $[\bullet]^+$ is an operation that sets negative values to zero. The parameter η is equal to 0 if $\sum_{k=1}^{K-1}[\hat{\mu}_k]^+ \leq 1$ and chosen such that $\sum_{k=1}^{K-1}\mu_k^{(\ell+1)} = 1$, otherwise. In contrast to the projected gradient descent method, which delivers a sequence of decreasing values, this method might yield an increase of the objective function in some iterations. Nevertheless, the method provably converges to the optimum if the step size is chosen such that [14]

$$\sum_{\ell=1}^{\infty}\alpha_\ell = \infty, \quad \alpha_\ell \to 0,\ \ell \to \infty.$$

This algorithm is summarized in Algorithm 3.12.

Algorithm 3.12 Projected subgradient method

$\mu_k^{(0)} \leftarrow \frac{1}{K},\ k = 1, \ldots, K-1$, $\ell \leftarrow 0$
Compute $g(\tilde{\boldsymbol{\mu}}^{(0)})$
repeat
  $\hat{\tilde{\boldsymbol{\mu}}}^{(\ell+1)} = \tilde{\boldsymbol{\mu}}^{(\ell)} - \alpha_\ell \boldsymbol{s}^{(\ell)}$
  $\mu_k^{(\ell+1)} = \left[\left[\hat{\mu}_k^{(\ell+1)}\right]^+ - \eta\right]^+, \quad k = 1, \ldots, K-1$
  $\ell \leftarrow \ell + 1$
  Compute $g(\tilde{\boldsymbol{\mu}}^{(\ell)})$
until $|g(\tilde{\boldsymbol{\mu}}^{(\ell)}) - g(\tilde{\boldsymbol{\mu}}^{(\ell-1)})| < \epsilon$

3.3.3 Implementation issues

Once a solution for Problem 3.43 has been found, it remains to compute the variables γ and $\boldsymbol{\rho}$ that achieve optimality in Problem 3.41, i.e., the primal problem. Due to strong duality $\bar{\gamma} = g(\bar{\tilde{\boldsymbol{\mu}}})$ where $\bar{\gamma}$ is the solution of the rate-balancing problem and $\bar{\tilde{\boldsymbol{\mu}}}$ the minimizer of the dual problem. Obviously, the optimum rate vector is given by $\bar{\boldsymbol{\rho}} = \bar{\gamma}\boldsymbol{q}$. However, it remains to find out what is the transmission strategy that leads to $\bar{\boldsymbol{\rho}}$. To this end, consider the following relationship (cf. Eqn A.8).

$$\bar{\gamma}(\bar{\boldsymbol{\rho}}) \leq \frac{\bar{R}_K}{q_K} + \sum_{k=1}^{K-1} \bar{\mu}_k \left(\frac{\bar{R}_k}{q_k} - \frac{\bar{R}_K}{q_K} \right) \leq g(\bar{\tilde{\boldsymbol{\mu}}}) = \bar{\gamma}(\bar{\boldsymbol{\rho}}).$$

Since for all inequalities, equality must hold, it becomes clear that $\bar{\boldsymbol{\rho}}$ is a maximizer of Eq. 3.44. That is, the optimum rate vector is a maximizer of the weighted sum-rate maximization problem with weights $w_k = \bar{\mu}_k/q_k$, $k = 1, \ldots, K-1$ and $w_K = (1 - \sum_{k=1}^{K-1} \bar{\mu}_k)/q_K$. If all these weights are different from each other, there is only a rate vector that maximizes the weighted sum-rate and the corresponding transmit statistics can be computed by solving Problem 3.22. This vector is $\bar{\boldsymbol{\rho}}$. The subgradient of $g(\tilde{\boldsymbol{\mu}})$ at $\bar{\tilde{\boldsymbol{\mu}}}$ is in this case zero. If some of the weights are equal, the optimum transmit covariance matrices in the MAC are still unique, but the optimum rate vectors are not (cf. Section 3.2). As discussed in Section 3.2, all these weighted sum-rate optimum rate vectors define a time-sharing region on the boundary of $R^{\mathrm{DPC}}(P)$ or, equivalently, $R^{\mathrm{MAC}}(P)$. The rate balancing optimum point is, in this case, just one of these vectors. If, for practical reasons, only successive encoding/decoding is considered, the problem consists in identifying the rate vectors achievable with successive encoding/decoding between which time-sharing should be performed in order to reach $\bar{\boldsymbol{\rho}}$. Besides, the time share corresponding to each of these vectors must be determined.

Assume that after solving the dual problem, M sets of users can be identified, all users of each set having identical weights. Let J_m be the cardinality of the mth set. Under this assumption, there is, in general, at least[4] $J = J_1!J_2!\cdots J_M!$ different rate vectors $\boldsymbol{\rho}_{1,\ldots,J}$ that can be achieved with successive decoding in the dual MAC and are weighted sum-rate maximizers. Each of these vectors corresponds to a different decoding order. As all these vectors lie on the same hyperplane of dimension $K-1$, Carathéodory's theorem on convex sets [40] can be invoked in order to show that the point of intersection of the constraint $\boldsymbol{\rho} = \gamma\boldsymbol{q}$ and this hyperplane lies in the convex hull of at most K of the J different vertices of the convex polytope defined by the weighted sum-rate maximizing rate vectors. That is, at most, time-sharing between K different orderings is required. If $J \leq K$ the time shares $\theta_{1,\ldots,J}$ corresponding to each of these vector can be computed by solving $\bar{\boldsymbol{\rho}} = \sum_{j=1}^{J} \boldsymbol{\rho}_j \theta_j$, which has a unique solution. On the contrary, if $J > K$ the resulting linear system is underdetermined and a solution has to be found that satisfies the constraints $\theta_j \geq 0$, i.e., no negative time shares are allowed, and $\sum_{j=1}^{J} \theta_j = 1$, i.e., the optimum is within the convex hull of the given rate vectors. A possible approach consists of sequentially considering each of the $\binom{J}{K}$ possible combinations of rate vectors until a combination is found that yields

[4] If the underlying transmission strategy is OFDM the number of possible rate vectors is larger as decoding order can be varied across carriers.

a linear system of equations such that the solution fulfils the constraints. In this case the time shares of all other non-selected rate vectors are set to zero. Alternatively, a feasible solution to the constrained linear system

$$\bar{\boldsymbol{\rho}} = \sum_{j=1}^{J} \boldsymbol{\rho}_j \theta_j,$$

subject to $\theta_j \geq 0$ and $\sum_{j=1}^{J} \theta_j = 1$ can be found by applying phase 1 of the simplex method [92].

The discussion above is somehow idealistic in that it assumes that the optimum of the dual problem is perfectly known. In practice, however, all what we have after a finite number of iterations is an estimate of the dual optimum and, based on this approximation, an estimate of the primal optimum must be computed. The general issue of estimating primal optima from approximate dual solutions is a current topic of research (see [85, 15] and references therein). For our particular problem, given the sequence of dual variables $\tilde{\boldsymbol{\mu}}^{(1)}, \ldots, \tilde{\boldsymbol{\mu}}^{(L)}$ obtained after L iterations, the sequence of primal variables $\gamma^{(1)}, \ldots, \gamma^{(L)}$, $\boldsymbol{\rho}^{(1)}, \ldots, \boldsymbol{\rho}^{(L)}$ can be considered, where $\boldsymbol{\rho}^{(\ell)}$ is a maximizer of Eq. 3.44 for $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}^{(\ell)}$ and

$$\gamma^{(\ell)} = \min_k \frac{R_k^{(\ell)}}{q_k}.$$

Let $\bar{\gamma}_L = \max\{\gamma^{(\ell)} | \ell = 1, \ldots, L\}$ and $\bar{g}_L = \min\{g(\tilde{\boldsymbol{\mu}}^{(\ell)}) | \ell = 1, \ldots, L\}$. If after L iterations

$$\bar{g}_L - \bar{\gamma}_L < \epsilon,$$

for a desired accuracy ϵ the search can be terminated. Any desired accuracy will always be reached if the optimum weights $w_{1,\ldots,K}$ are all unequal (see Fig. 3.1). Unfortunately, if the primal optimum rate vector lies on a time-sharing region a gap will remain between $\bar{g}$ and $\bar{\gamma}$ no matter how many iterations are carried out (see Fig. 3.2). That is, even if convergence is achieved in the dual, in the primal, no convergence is reached. This is due to the fact that in Eq. 3.44 only maximizers $\boldsymbol{\rho}^{(\ell)}$ are considered that are achievable with successive decoding. These vectors will, in general, not be maximizers of the primal. In this case, let $\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(L)}$ be the sequence of weight vectors with $w_k^{(\ell)} = \mu_k^{(\ell)}/q_k$, $k = 1, \ldots, K$. Recalling that the relative order of the entries of these vectors indicates the order in which the users are optimally decoded, we can check where the last changes in the decoding order of the users occur in this sequence (cf. [50]). Assume that C changes in the decoding order are found in the last elements[5] of the sequence $\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(L)}$ and let $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_C$ be the rate vectors corresponding to those weight vectors that represent a change in the decoding order with respect to a previous weight vector. A good approximation of the primal solution can

[5]The last elements of the sequence can be defined as those in which the entries of the vectors $\boldsymbol{w}^{(\ell)}$ do not significantly change.

be found by solving

$$\max_{\theta_{1,\ldots,C}} \gamma \quad \text{subject to} \quad \gamma \boldsymbol{q} = \sum_{c=1}^{C} \boldsymbol{\rho}_c \theta_c,$$
$$\sum_{c=1}^{C} \theta_c = 1, \quad \theta_c \geq 0, \; c = 1, \ldots, C.$$

Beside these subtleties concerning the practical implementation, a general problem of subgradient-based optimum approaches is the relatively slow convergence rate, which appears to be sensitive to the amount of users in the system (see Fig. 3.3). In addition, each iteration involves execution of an iterative weighted sum-rate algorithm. Although these algorithms have been empirically observed to have a good convergence behavior [62], their iterative character represents a further source of non-deterministic complexity in the computation of the optimum rate-balancing solution. These facts motivate the introduction and discussion of simple non-iterative approaches in the next chapter that can achieve a large fraction of the performance achieved by optimum iterative approaches with a complexity which can be determined beforehand independently of particular system parameters.

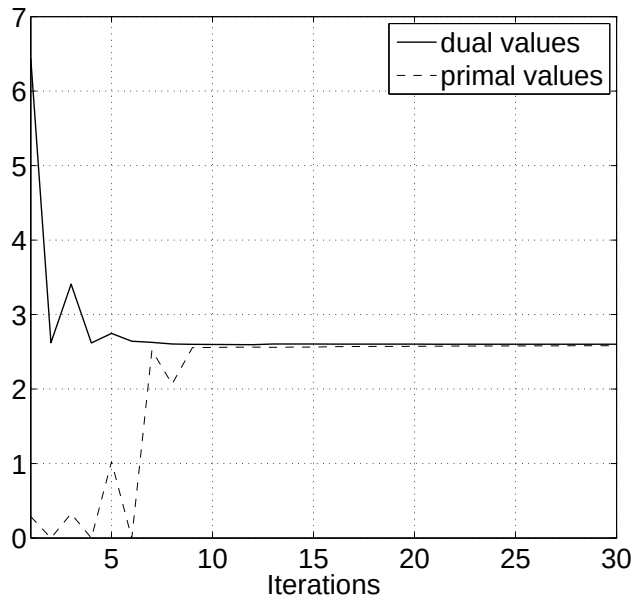

Figure 3.1: Dual values $g^{(\ell)}$ and corresponding primal values $\gamma^{(\ell)}$ during the first 30 iterations of the ellipsoid algorithm for a MIMO OFDM broadcast channel with $N = 16$, $K = 3$, $t = 4$ and $r_k = 2$, $\forall k$, SNR = 20 dB. The vector of relative rates is given by $\boldsymbol{q} = [1, 3, 6]^{\mathrm{T}}$ and the optimum weights $\boldsymbol{w} = [0.0214, 0.0400, 0.1431]^{\mathrm{T}}$, i.e., no time-sharing is required to achieve the rate-balancing solution.

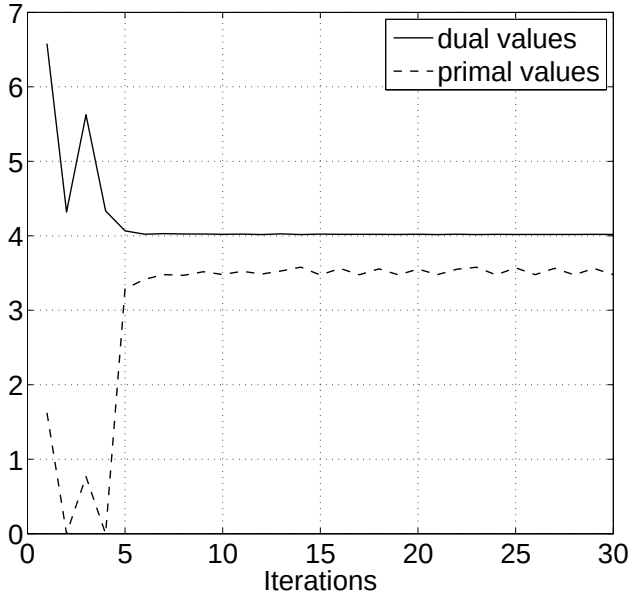

Figure 3.2: Dual values $g^{(\ell)}$ and corresponding primal values $\gamma^{(\ell)}$ during the first 30 iterations of the ellipsoid algorithm for a MIMO OFDM broadcast channel with $N = 16$, $K = 3$, $t = 4$, $r_k = 2$, $\forall k$, SNR $= 20$ dB. The vector of relative rates is given by $\boldsymbol{q} = [1, 3, 3]^{\mathrm{T}}$ and the optimum weights $\boldsymbol{w} = [0.0775, 0.1537, 0.1537]^{\mathrm{T}}$, i.e., time-sharing between users 2 and 3 is required to achieve the rate balancing solution.

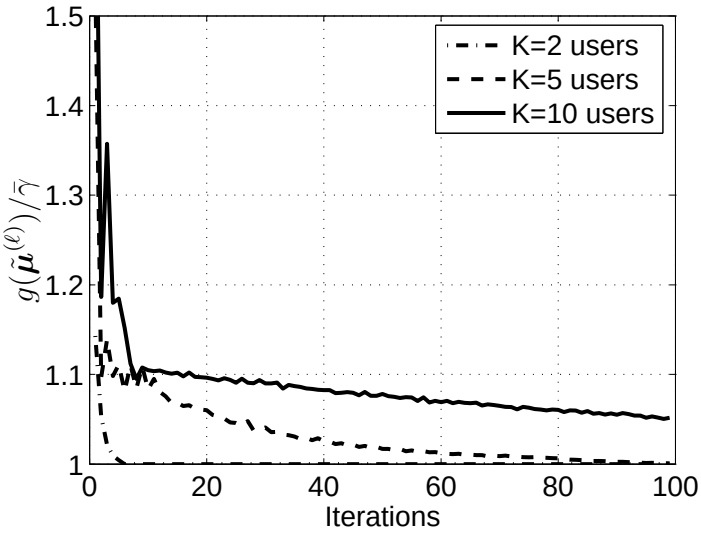

Figure 3.3: Convergence of the ellipsoid method applied to the dual of the rate-balancing problem. Averaged curves over 100 channel realizations. $N = 16$, $t = 4$, $r_k = 2$, SNR $= 10$ dB, $\boldsymbol{q} = [1, \ldots, 1]^{\mathrm{T}}$.

4 Non-iterative approaches for the broadcast channel

Finding the optimum solutions of the three problems discussed in the previous chapter requires the application of iterative algorithms. While for the weighted sum-rate maximization problem, and the sum-rate maximization problem as an especial case, recently proposed algorithms are observed to require few iterations to reach convergence, for the rate balancing problem, existing subgradient-based algorithms exhibit a poor convergence behavior. In any case, for any of the algorithms described in the previous chapter, the number of iterations required in order to achieve convergence is a function of the system parameters that can not be predicted beforehand and, therefore, constitutes a source of uncertainty regarding the computational complexity needed to compute the optimum solution. In some applications where delay is an issue, such as wireless communications with fast time-varying channels, this feature of optimum approaches might become a problem.

In this chapter, a number of suboptimum approaches are discussed that deliver solutions that require a closed number of computations. Most of the chapter deals with decomposition approaches that transform the original broadcast channel into a set of scalar decoupled subchannels. The resulting subchannels are decoupled in the sense that transmission in any subchannel does not cause interference on all other subchannels. First, we present a general framework that encloses all decomposition approaches discussed in the literature so far and a general successive subchannel allocation method that can be applied to both linear schemes and schemes based on successive encoding. Then, we consider optimization of the subchannel and power allocation policies for each of the design problems discussed in the previous chapter. In the last section of the chapter, a novel successive subchannel allocation approach based on successive encoding is introduced that allows for cross-talk between the resulting scalar subchannels.

4.1 Broadcast channel decomposition schemes

Without loss of optimality the transmit signal for the general Gaussian broadcast channel given by Eq. 2.6 can be written as

$$\boldsymbol{x} = \sum_{k=1}^{K} \boldsymbol{V}_k \boldsymbol{P}_k^{1/2} \boldsymbol{s}_k,$$

where $\boldsymbol{V}_k \in \mathbb{C}^{t \times m_k}$ is a matrix with orthonormal column vectors, $\boldsymbol{P}_k \in \mathbb{R}^{m_k \times m_k}$ is a diagonal power matrix and $\boldsymbol{s}_k \in \mathbb{C}^{m_k \times 1}$ is the vector of signals intended for user k, which is assumed to be a realization of a zero-mean, circularly symmetric complex Gaussian distributed vector $\mathbf{s}_k$ with covariance $\mathrm{E}\{\mathbf{s}_k \mathbf{s}_k^{\mathrm{H}}\} = \boldsymbol{I}_{m_k}$. Signals intended for different users

are assumed to be statistically independent. The number of spatial dimensions m_k is less than or equal to $\min\{t, r_k\}$ and

$$\sum_{k=1}^{K} \mathrm{Tr}\{\boldsymbol{P}_k\} \leq P.$$

That this structure of the transmit signal is optimum is a simple consequence of the fact that every transmit covariance matrix admits an eigenvalue decomposition. There also exist matrices of orthonormal columns $\boldsymbol{U}_k \in \mathbb{C}^{r_k \times m_k}$, $k = 1, \ldots, K$, that can be applied at the receivers without capacity loss, i.e., $I(\boldsymbol{s}_k, \boldsymbol{y}_k) = I(\boldsymbol{s}_k, \boldsymbol{U}_k^{\mathrm{H}} \boldsymbol{y}_k)$. For any user k, one such matrix is the matrix of right singular vectors of the corresponding matched filter matrix (cf. [115]).

In this section, we are concerned with the choice of precoding matrices $\boldsymbol{V}_k$ and receive filter matrices $\boldsymbol{U}_k$. However, rather than optimality our goal is simplicity. In particular, we aim at finding precoding and receive filter matrices that decompose the broadcast channel into a set of decoupled subchannels. This can easily be done by using elementary tools such as zero-forcing constraints or the singular value decomposition (SVD). Obviously, there is not a unique way of decomposing a broadcast channel. This degree of freedom can be exploited in order to optimize the performance measure of interest. A further optimization step can be carried out subsequently by choosing an adequate power allocation policy for the set of resulting non-interfering subchannels. In the following, we distinguish between linear decomposition approaches and successive-encoding-based decompostion approaches. In the latter, part of the interference is eliminated by resorting to dirty paper coding. In the former, the choice of beamforming vectors is made such that interference is completely suppressed.

4.1.1 Linear decomposition

A linear decomposition of the broadcast channel is achieved if the following two conditions hold

$$\boldsymbol{U}_k^{\mathrm{H}} \boldsymbol{H}_k \boldsymbol{V}_k = \boldsymbol{D}_k, \quad k = 1, \ldots, K, \tag{4.1}$$

$$\boldsymbol{U}_j^{\mathrm{H}} \boldsymbol{H}_j \boldsymbol{V}_k = \boldsymbol{0}, \quad k = 1, \ldots, K, \ \forall j \neq k, \tag{4.2}$$

where $\boldsymbol{D}_k \in \mathbb{R}_+^{m_k \times m_k}$ is diagonal. That is, information for any user k is transmitted over a set of m_k decoupled channels and the signals transmitted to this user do not cause interference to signals received by any other user j. The first condition is not restrictive in terms of achievable capacity. In fact, for any matrix $\tilde{\boldsymbol{V}}_k$ of orthonormal columns that satisfies the second condition and a matrix $\tilde{\boldsymbol{U}}_k$ of orthonormal columns that filters out interference from other users, we can compute the SVD of the product $\tilde{\boldsymbol{U}}_k^{\mathrm{H}} \boldsymbol{H}_k \tilde{\boldsymbol{V}}_k$. Let $\bar{\boldsymbol{U}}_k \in \mathbb{C}^{m_k \times m_k}$ be the resulting unitary matrix of left singular vectors and $\bar{\boldsymbol{V}}_k \in \mathbb{C}^{m_k \times m_k}$ the corresponding unitary matrix of right singular vectors. The matrix $\boldsymbol{V}_k = \tilde{\boldsymbol{V}}_k \bar{\boldsymbol{V}}_k$ has orthonormal columns and satisfies the second condition. The matrix $\boldsymbol{U}_k = \tilde{\boldsymbol{U}}_k \bar{\boldsymbol{U}}_k$ has orthonormal columns and suppresses interference from all other users. Together, $\boldsymbol{U}_k$ and $\boldsymbol{V}_k$ satisfy the first condition and the resulting channel matrix $\boldsymbol{D}_k$ supports the same transmission rate as the initial $\tilde{\boldsymbol{U}}_k^{\mathrm{H}} \boldsymbol{H}_k \tilde{\boldsymbol{V}}_k$ matrix. The second condition demands that the

number of subchannels in the system be non-larger than the number of transmit antennas, i.e., $\sum_{k=1}^{K} m_k \leq t$.

Given a performance measure of interest, optimization can be carried out over the choice of m_k, $k = 1, \ldots, K$. But even fixing the number of subchannels per user, in general, there are additional degrees of freedom in the choice of precoding and receive filter matrices. In order to see this, consider arbitrarily chosen receive filter matrices $\tilde{\boldsymbol{U}}_k$ of orthonormal columns and dimension $r_k \times m_k$, $k = 1, \ldots, K$, where $\sum_{k=1}^{K} m_k \leq t$. For each user k, consider the matrix

$$\bar{\boldsymbol{H}}_k = [\ \boldsymbol{H}_1^{\mathrm{T}}\tilde{\boldsymbol{U}}_1^* \ \cdots \ \boldsymbol{H}_{k-1}^{\mathrm{T}}\tilde{\boldsymbol{U}}_{k-1}^* \ \boldsymbol{H}_{k+1}^{\mathrm{T}}\tilde{\boldsymbol{U}}_{k+1}^* \ \cdots \ \boldsymbol{H}_K^{\mathrm{T}}\tilde{\boldsymbol{U}}_K^* \ ]^{\mathrm{T}}, \tag{4.3}$$

and the projector associated with the null space of this matrix $\boldsymbol{T}_k^{\perp} = \boldsymbol{I}_t - \bar{\boldsymbol{H}}_k^{\mathrm{H}}(\bar{\boldsymbol{H}}_k\bar{\boldsymbol{H}}_k^{\mathrm{H}})^{-1}\bar{\boldsymbol{H}}_k$. Let $\bar{\boldsymbol{U}}_k \in \mathbb{C}^{m_k \times m_k}$ be the matrix of m_k dominant left singular vectors and $\boldsymbol{V}_k \in \mathbb{C}^{t \times m_k}$ the corresponding matrix of right singular vectors obtained from performing a SVD on the product $\tilde{\boldsymbol{U}}_k^{\mathrm{H}}\boldsymbol{H}_k\boldsymbol{T}_k^{\perp}$. Obviously, $\boldsymbol{T}_k^{\perp}\boldsymbol{V}_k = \boldsymbol{V}_k$. Using this and the property $\bar{\boldsymbol{H}}_k\boldsymbol{T}_k^{\perp} = \boldsymbol{0}$, it can easily be shown that $\boldsymbol{V}_k$ satisfies Eq. 4.2 with $\boldsymbol{U}_j = \tilde{\boldsymbol{U}}_j\bar{\boldsymbol{U}}_j$, $\forall j \neq k$. Further, $\boldsymbol{U}_k = \tilde{\boldsymbol{U}}_k\bar{\boldsymbol{U}}_k$ and $\boldsymbol{V}_k$ satisfy Eq. 4.1. That is, starting from any arbitrary choice of matrices $\tilde{\boldsymbol{U}}_{1,\ldots,K}$ a set of precoding matrices and receive filter matrices can be found. In general, the resulting subchannels, whose gains are given by the diagonal entries of $\boldsymbol{D}_{1,\ldots,K}$, will be different for different choices of matrices $\tilde{\boldsymbol{U}}_{1,\ldots,K}$. Optimization of the number of subchannels for each user and the choice of precoding and receiver filter matrices can be viewed as a general subchannel allocation problem. This is in general difficult to solve even for very elemental performance measures such as sum rate mostly due to its combinatorial nature. However, a simple suboptimum subchannel allocation scheme will be presented in Section 4.1.3.1 that, as we shall see, delivers very good performance. Before that, in the following, we review some of the known linear decomposition schemes and existing partial solutions to this subchannel allocation problem.

If $r_1 = \cdots = r_K = 1$, the subchannel allocation problem reduces to a user selection problem. The objective is to select a group of users $\mathcal{U}$ of cardinality $|\mathcal{U}| \leq t$ such that the performance measure of interest is maximized.[1] Let $U \leq t$ be the number of selected users and $\boldsymbol{h}_i \in \mathbb{C}^{t \times 1}$, $i = 1, \ldots, U$, their corresponding vector channels. The beamforming transmit vector for user i can optimally[2] be computed as $\boldsymbol{v}_i = \boldsymbol{T}_i^{\perp}\boldsymbol{h}_i / \|\boldsymbol{T}_i^{\perp}\boldsymbol{h}_i\|_2$, where $\boldsymbol{T}_i^{\perp} = \boldsymbol{I}_t - \bar{\boldsymbol{H}}_i^{\mathrm{H}}(\bar{\boldsymbol{H}}_i\bar{\boldsymbol{H}}_i^{\mathrm{H}})^{-1}\bar{\boldsymbol{H}}_i$ and[3]

$$\bar{\boldsymbol{H}}_i = [\ \boldsymbol{h}_1 \ \cdots \ \boldsymbol{h}_{i-1} \ \boldsymbol{h}_{i+1} \ \cdots \ \boldsymbol{h}_U \ ]^{\mathrm{H}}.$$

Equivalently, if we define $\boldsymbol{H}(\mathcal{U}) \in \mathbb{C}^{t \times U}$ as the matrix whose column i is given by $\boldsymbol{h}_i$, the transmit beamforming vectors can be obtained by computing the scaled Moore-Penrose pseudoinverse of $\boldsymbol{H}(\mathcal{U})^{\mathrm{H}}$ as

$$\boldsymbol{V} = \boldsymbol{H}(\mathcal{U})(\boldsymbol{H}(\mathcal{U})^{\mathrm{H}}\boldsymbol{H}(\mathcal{U}))^{-1}\boldsymbol{S},$$

[1] If QoS constraints are considered as in the rate-balancing optimization problem, the feasible set might be empty, for instance, if $K > t$ and $q_k > 0$, $\forall k$ (cf. Section 3.3). In order to circumvent this problem, subchannel allocation on different subcarriers or time-slots should be considered.

[2] Other choices of $\boldsymbol{v}_i$ are possible if $U < t$, however the resulting channel gains are smaller.

[3] For single-antenna receivers the channel matrix becomes a row vector. Since in this work vectors are column vectors, the channel matrix of a single-antenna user is denoted by a conjugate transposed column vector. In particular, for user k, $\boldsymbol{H}_k = \boldsymbol{h}_k^{\mathrm{H}}$.

where the ith column of $\boldsymbol{V}$ is $\boldsymbol{v}_i$ and $\boldsymbol{S}$ is a diagonal scaling matrix with entries

$$s_{i,i} = \left(\sqrt{[(\boldsymbol{H}(\mathcal{U})^{\mathrm{H}} \boldsymbol{H}(\mathcal{U}))^{-1}]_{i,i}} \right)^{-1}.$$

This scheme is commonly known as zero-forcing beamforming and has been considered in numerous recent publications, e.g., [42, 139, 22, 18, 110]. Several works have addressed the problem of user selection in the context of sum-rate maximization [42, 139], where asymptotic optimality of zero-forcing has been shown for a large number of users [139]. In [110] this problem has also been discussed in a rate-balancing context.

For the case of multiple receive antennas, decomposition algorithms were independently proposed in [107, 31, 93]. In all these works, no decision is made a priori regarding the number of dimensions assigned to the users. Instead, using the notation employed in Eq. 4.3, the receive filter matrices are initially chosen to be $\tilde{\boldsymbol{U}}_k = \boldsymbol{I}_{r_k}$, $k = 1, \ldots, K$. Subsequently, after computing the matrices $\bar{\boldsymbol{H}}_k$ and $\boldsymbol{T}_k^{\perp}$ for each user, a SVD of the products $\boldsymbol{H}_k \boldsymbol{T}_k^{\perp}$ is performed. The resulting matrices of left singular vectors $\boldsymbol{U}_k \in \mathbb{C}^{r_k \times m_k}$ and right singular vectors $\boldsymbol{V}_k \in \mathbb{C}^{t \times m_k}$ satisfy Eqs. 4.2 and 4.1. For each user, the number of dimensions m_k corresponds to the number of non-zero singular values. A necessary condition for the applicability of this scheme is

$$t > \sum_{j \neq k} r_j, \quad k = 1, \ldots, K. \tag{4.4}$$

In order to make the applicability of this algorithm possible when this condition is violated, user selection schemes have been proposed in [104, 49, 106] among others. Fixing a priori the number of subchannels m_k that each user should get allocated, an iterative algorithm has been proposed in [30] in order to optimize the choice of receive filter and precoding matrices. The algorithm starts with matrices $\tilde{\boldsymbol{U}}_k^{(0)} \in \mathbb{C}^{r_k \times m_k}$ being the matrices of m_k dominant left singular vectors of matrices $\boldsymbol{H}_k$. Then, after computing the matrices $\bar{\boldsymbol{H}}_k$ and $\boldsymbol{T}_k^{\perp}$ for each user, a SVD of the products $\tilde{\boldsymbol{U}}_k^{(0),\mathrm{H}} \boldsymbol{H}_k \boldsymbol{T}_k^{\perp}$ is performed. Denoting with $\bar{\boldsymbol{U}}_k^{(0)}$ the resulting matrices of left singular vectors, the new receive filter matrices are given by $\tilde{\boldsymbol{U}}_k^{(1)} = \tilde{\boldsymbol{U}}_k^{(0)} \bar{\boldsymbol{U}}_k^{(0)}$. The same computations are repeated using these new matrices and so on until convergence is reached. While no analytical proof of convergence is provided, empirical evidence suggests that this algorithm has a reliable convergence behavior. The problem of choosing the number of subchannels assigned to each user has been addressed in two different forms. The authors in [106, 139] propose to use the matrices of left singular vectors $\boldsymbol{U}_k$ of the channel matrices $\boldsymbol{H}_k$ as receive filter matrices and treat the rows of $\boldsymbol{U}_k^{\mathrm{H}} \boldsymbol{H}_k$ as non-cooperative channels. This allows to employ user grouping algorithms proposed for the MISO setting in order to solve the subchannel allocation problem. In [137], a greedy antenna selection algorithm has been proposed based on a sum-rate criterion. This short overview reveals that the different aspects of this general subchannel allocation problem for linear decomposition schemes have been separately treated following somehow disconnected approaches. In Section 4.1.3.1 a general and compact successive allocation scheme is presented that allows for a joint optimization of precoding and receive filter matrices and the number of subchannels to be allocated to each user.

4.1.2 Successive-encoding-based decomposition

If the broadcast channel is linearly decomposed, the precoding vectors of any user k must satisfy $\sum_{j\neq k} m_j$ orthogonality constraints. A way of increasing the number of degrees of freedom in the choice of precoding vectors consists of performing a successive encoding of the information streams transmitted over the assigned subchannels and applying a dirty paper coding scheme to each stream in order to fully neutralize interference caused by previously encoded streams. In this way, the precoding vector corresponding to a certain subchannel must only satisfy orthogonality constraints with respect to the subchannels whose information streams are encoded at an earlier stage but no constraints are imposed by those subchannels whose information streams are encoded later. The resulting subchannels are physically coupled in the sense that any subchannel causes interference to the subchannels where information is subsequently encoded. However, this interference does not have any impact on the achievable information rate over these subchannels and, therefore, we can say that subchannels are virtually decoupled.

Let $\pi : \{1,\ldots,t\} \rightarrow \{1,\ldots,K\}$ be a subchannel allocation function that assigns a subchannel to the user to which this subchannel belongs. Also, let the domain of this function indicate the order in which the subchannels are encoded. A virtual decomposition of the broadcast channel is achieved if

$$\boldsymbol{u}_i^{\mathrm{H}} \boldsymbol{H}_{\pi(i)} \boldsymbol{v}_j = 0, \quad j > i,\ i,j \in \{1,\ldots,t\}\,.$$

Here, $\boldsymbol{v}_i \in \mathbb{C}^{t\times 1}$ and $\boldsymbol{u}_i \in \mathbb{C}^{r_k\times 1}$ are, respectively, the unit-norm transmit and receive beamforming vectors corresponding to subchannel i. Note that, as in the linear approaches, a maximum of t dimensions can be allocated as no more than $t-1$ orthogonality constraints can be satisfied in a space of dimension t. Beside the number of dimensions allocated to each user and the particular choice of transmit and receive beamforming vectors, the choice of encoding order is a further degree of freedom that can be exploited in order to optimize a performance measure of interest.

For $r_1 = \cdots = r_K = 1$ and a given subchannel allocation function, the transmit beamforming vectors can be optimally computed as $\boldsymbol{v}_i = \boldsymbol{T}_i^{\perp}\boldsymbol{h}_{\pi(i)}/\|\boldsymbol{T}_i^{\perp}\boldsymbol{h}_{\pi(i)}\|_2$, where $\boldsymbol{T}_i^{\perp} = \boldsymbol{I}_t - \bar{\boldsymbol{H}}_i^{\mathrm{H}}(\bar{\boldsymbol{H}}_i\bar{\boldsymbol{H}}_i^{\mathrm{H}})^{-1}\bar{\boldsymbol{H}}_i$ and

$$\bar{\boldsymbol{H}}_i = [\ \boldsymbol{h}_{\pi(1)} \quad \cdots \quad \boldsymbol{h}_{\pi(i-1)}\ ]^{\mathrm{H}}.$$

Equivalently, if we define $\boldsymbol{H}(\pi) \in \mathbb{C}^{t\times U}$ as the matrix whose column i is given by $\boldsymbol{h}_{\pi(i)}$ with $U = \min\{t,K\}$, the transmit beamforming vectors can be obtained by performing a QR factorization of $\boldsymbol{H}(\pi)$. That is, the transmit beamforming vector $\boldsymbol{v}_i$ is the ith column of $\boldsymbol{V}$, where $\boldsymbol{H}(\pi) = \boldsymbol{V}\boldsymbol{R}$ and $\boldsymbol{R}$ is an upper triangular matrix whose entries on the main diagonal represent the channel gains of the resulting subchannels. This approach is commonly known as zero-forcing dirty-paper coding and has been discussed in [23, 97, 123] and previously in combination with Tomlinson-Harashima precoding in [56]. Since, for a given number of allocated subchannels, the allocation of additional subchannels encoded at a later stage is not detrimental, $U = \min\{t,K\}$ dimensions can always be allocated without loss of optimality. This is in contrast to the zero-forcing beamforming approach, for which $U < \min\{t,K\}$ may yield a better performance in some cases. If $K \leq t$, optimization of

any performance measure of interest can be performed over the choice of encoding order. If $K > t$, both user selection and encoding order are the degrees of freedom that can be exploited in order to optimize performance. In [136, 123] greedy algorithms are presented that optimize encoding order using sum-rate as a figure of merit. While the algorithm in [136] is only applicable to systems with $K \leq t$, the algorithm in [123] can be applied to systems with $K > t$. In that case, the algorithm performs user selection and optimization of the encoding order at the same time.

For multiple receive antennas, decomposition algorithms have been presented in [108, 41, 79]. In both [108, 41] a block-wise decomposition approach is proposed along the lines of the linear decomposition approaches presented in [107, 31, 93]. Assuming that the streams of information for a certain user are encoded in the order indicated by the user index, i.e., streams of information for user 1 are encoded in first place, those for user 2 in the second place and so on, the transmit and receive beamforming vectors for user k are obtained as follows. First,

$$\bar{\boldsymbol{H}}_k = [\ \boldsymbol{H}_1^{\mathrm{T}} \quad \cdots \quad \boldsymbol{H}_{k-1}^{\mathrm{T}}\ ]^{\mathrm{T}}$$

is defined and the null-space projector of this matrix $\boldsymbol{T}_k^{\perp} = \boldsymbol{I}_t - \bar{\boldsymbol{H}}_k^{\mathrm{H}}(\bar{\boldsymbol{H}}_k\bar{\boldsymbol{H}}_k^{\mathrm{H}})^{-1}\bar{\boldsymbol{H}}_k$ is computed. Then, a SVD of the product $\boldsymbol{H}_k\boldsymbol{T}_k^{\perp}$ is performed. The matrix $\boldsymbol{V}_k$ of right singular vectors becomes the precoding matrix for user k and the matrix $\boldsymbol{U}_k$ of corresponding left singular vectors becomes the receive filter matrix for user k. Since the columns of $\boldsymbol{V}_k$ lie in the subspace associated with $\boldsymbol{T}_k^{\perp}$, $\boldsymbol{V}_k = \boldsymbol{T}_k^{\perp}\boldsymbol{V}_k$ holds, and, as a result, $\boldsymbol{U}_k$ and $\boldsymbol{V}_k$ diagonalize $\boldsymbol{H}_k$. That is, the subchannels corresponding to a particular user are completely decoupled and, hence, they can be independently encoded. The subchannels of user k do not cause interference on subchannels of users $1, \ldots, k-1$. By contrast, these subchannels cause interference on the subchannels of user k, which can be neutralized by using dirty paper coding. For a given group of users satisfying Eq. 4.4, a heuristic encoding order is proposed in [108]. Important issues such as user selection and optimization of the number of dimensions allocated to each user are not considered in these works. In [79] a successive allocation algorithm is proposed where each user, if served, is allocated just one spatial dimension. At each step, the user is selected that can transmit over a subchannel with maximum channel gain among those that satisfy orthogonality constraints with previously established subchannels. Correspondingly, the encoding order coincides with the order in which subchannels are allocated. This algorithm has been shown to have an optimum behavior in terms of sum-rate for large number of users.

In the following section an algorithm is presented that comprises all these state-of-the-art decomposition approaches as particular cases and incorporates mechanisms in order to perform user selection, optimization of the number of dimensions allocated to each user and optimization of the encoding order for any performance measure of interest. Different aspects of this general decomposition algorithm have been discussed in [113, 115, 118, 115, 16, 116, 114].

4.1.3 Successive subchannel allocation method

The decomposition algorithm that we present in this section proceeds successively assigning a new spatial dimension to a particular user in each step. The set of eligible dimensions

in each step is given by the set of singular values of all users in the network within the orthogonal subspace to that spanned by previously assigned dimensions. Constraining the set of candidate subchannels to be orthogonal to the set of established subchannels, we make sure that the dimension allocated in a certain step does not cause interference on previously assigned subchannels. Choosing the encoding order to be the same as the order in which subchannels are allocated makes possible to neutralize the interference caused by previously established subchannels on new ones by means of coding. Selection of the spatial dimension that is assigned at a given step is made according to a rule that can be defined in accordance with some performance measure of interest. In this way, the users served, the number of dimensions assigned to each user, the transmit and receive beamforming vectors corresponding to a certain dimension and the encoding order become all parameters that are implicitly determined by the performance measure of interest via the associated channel allocation rule.

Specifically, the algorithm works as follows. After having established the first $j-1$ spatial subchannels, the projection matrix $\boldsymbol{T}_j$ is computed as

$$\boldsymbol{T}_j = \boldsymbol{T}_{j-1} - \boldsymbol{v}_{j-1}\boldsymbol{v}_{j-1}^{\mathrm{H}},$$

where $\boldsymbol{T}_1 = \boldsymbol{I}_t$ and $\boldsymbol{v}_{j-1}$ is the transmit beamforming vector corresponding to the dimension just allocated in the previous step. As it will become clear later, matrix $\boldsymbol{T}_j$ represents the projector of the subspace defined by the intersection of the kernels of the $j-1$ previously established subchannels. Then, channel matrices of all users are projected into this subspace,

$$\boldsymbol{H}_k^j = \boldsymbol{H}_k\boldsymbol{T}_j, \quad k = 1, \ldots, K,$$

and singular value decompositions of all projected channel matrices are performed,

$$\boldsymbol{H}_k^j = \boldsymbol{U}_k^j\boldsymbol{\Lambda}_k^j\boldsymbol{V}_k^{j,\mathrm{H}}, \quad k = 1, \ldots, K.$$

At this stage, among the set of potential subchannels one is selected according to any particular rule. Denoting by $\mathcal{R}$ the rule that selects one out of all possible subchannels we can mathematically write

$$(\bar{k}, \bar{s}) = \mathcal{R}\left(\left\{\lambda_{k,s}^j | k = 1, \ldots, K,\ s = 1, \ldots, \rho_k^j\right\}\right), \quad \pi(j) = \bar{k},$$

$$\boldsymbol{v}_j = \boldsymbol{V}_{\bar{k}}^j\boldsymbol{e}_{\bar{s}}, \quad \boldsymbol{u}_j = \boldsymbol{U}_{\bar{k}}^j\boldsymbol{e}_{\bar{s}},$$

where $\lambda_{k,s}^j$ is the sth eigenvalue in the main diagonal of matrix $\boldsymbol{\Lambda}_k^j$, $\rho_k^j = \mathrm{Rank}\left\{\boldsymbol{H}_k^j\right\}$ and $\boldsymbol{e}_s$ is a column vector with a 1 in the sth row and zeros elsewhere. The rule $\mathcal{R}$ can be viewed as a function that takes the set of singular values in the remaining spatial subspace and returns an ordered pair of indexes that identify the selected dimension. Internally, $\mathcal{R}$ might also make use of further system parameters such as quality-of-service constraints or scheduling statistics. In order to allocate the $(j+1)$th spatial subchannel the same procedure is repeated. This allocation method is summarized in Algorithm 4.1. For convenience, henceforth, we will refer to this algorithm as successive encoding successive allocation method (SESAM).

Algorithm 4.1 Successive encoding successive subchannel allocation method

1: $j \leftarrow 1, \quad \boldsymbol{T}_1 \leftarrow \boldsymbol{I}_t$

2: **repeat**

3: $\quad \boldsymbol{H}_k^j \leftarrow \boldsymbol{H}_k \boldsymbol{T}_j, \quad k = 1, \ldots, K$

4: $\quad \boldsymbol{H}_k^j \leftarrow \boldsymbol{U}_k^j \boldsymbol{\Lambda}_k^j \boldsymbol{V}_k^{j,\mathrm{H}}, \quad k = 1, \ldots, K$

5: $\quad (\bar{k}, \bar{s}) \leftarrow \mathcal{R}\left(\left\{\lambda_{k,s}^j | k = 1, \ldots, K,\ s = 1, \ldots, \rho_k^j\right\}\right),\ \pi(j) \leftarrow \bar{k},$

$\quad \boldsymbol{v}_j \leftarrow \boldsymbol{V}_{\bar{k}}^j \boldsymbol{e}_{\bar{s}},\ \boldsymbol{u}_j \leftarrow \boldsymbol{U}_{\bar{k}}^j \boldsymbol{e}_{\bar{s}}$

6: $\quad \boldsymbol{T}_{j+1} \leftarrow \boldsymbol{T}_j - \boldsymbol{v}_j \boldsymbol{v}_j^{\mathrm{H}}, \quad j \leftarrow j + 1$

7: **until** $j > \sum_k r_k$ or $\boldsymbol{T}_j = \boldsymbol{0}$

For subchannel j, interference caused by subchannels $i > j$ is forced to zero, i.e.,

$$\boldsymbol{u}_j^{\mathrm{H}} \boldsymbol{H}_{\pi(j)} \boldsymbol{v}_{i>j} = 0.$$

In order to see this consider the following equations,

$$\begin{aligned} \boldsymbol{u}_j^{\mathrm{H}} \boldsymbol{H}_{\pi(j)} \boldsymbol{v}_{i>j} &= \\ \boldsymbol{u}_j^{\mathrm{H}} \boldsymbol{H}_{\pi(j)} \boldsymbol{T}_{i>j} \boldsymbol{v}_{i>j} &= && (4.5) \\ \boldsymbol{u}_j^{\mathrm{H}} \boldsymbol{H}_{\pi(j)} \boldsymbol{T}_j \boldsymbol{T}_{i>j} \boldsymbol{v}_{i>j} &= && (4.6) \\ \lambda_{\bar{k},\bar{s}}^j \boldsymbol{v}_j^{\mathrm{H}} \boldsymbol{T}_{i>j} \boldsymbol{v}_{i>j} &= 0. && (4.7) \end{aligned}$$

In Eq. 4.5, we make use of the fact that $\boldsymbol{v}_{(i>j)}$ lies within the subspace spanned by $\boldsymbol{T}_{i>j}$. In Eq. 4.6, we consider the fact that the image of $\boldsymbol{T}_{i>j}$ is within the subspace spanned by $\boldsymbol{T}_j$. Finally, in Eq. 4.7, we note that $\boldsymbol{u}_j$ is a left singular vector of $\boldsymbol{H}_{\pi(j)} \boldsymbol{T}_j$ with $\boldsymbol{v}_j$ as corresponding right singular vector, which, by construction, happens to be perpendicular to the subspace spanned by $\boldsymbol{T}_{i>j}$. By contrast, interference caused by subchannels $i < j$ is, in general, not eliminated by the choice of beamforming vectors. Note that in this case $\boldsymbol{T}_j \boldsymbol{T}_{i<j} = \boldsymbol{T}_j$ and, therefore, Eq. 4.6 does not hold. This interference can be neutralized by coding. An exception occurs when $\pi(i) = \pi(j)$ with $i \neq j$, i.e., when a same user gets allocated two different subchannels. In such case, it can be shown that subchannels j and i are entirely decoupled as follows. Assume $i > j$, then,

$$\begin{aligned} 0 &= \boldsymbol{u}_j^{\mathrm{H}} \boldsymbol{H}_{\pi(j)} \boldsymbol{v}_i \\ &= \boldsymbol{u}_j^{\mathrm{H}} \boldsymbol{H}_{\pi(j)} \boldsymbol{T}_i \boldsymbol{v}_i \\ &= \boldsymbol{u}_j^{\mathrm{H}} \boldsymbol{u}_i \lambda_{\bar{k},\bar{s}}^i, \end{aligned}$$

which shows that $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$ are necessarily orthogonal. On the other hand, interference

caused by subchannel j on subchannel i is given by

$$\begin{aligned} \boldsymbol{u}_i^{\mathrm{H}} \boldsymbol{H}_{\pi(i)} \boldsymbol{v}_j &= \\ \boldsymbol{u}_i^{\mathrm{H}} \boldsymbol{H}_{\pi(i)} \boldsymbol{T}_j \boldsymbol{v}_j &= \\ \boldsymbol{u}_i^{\mathrm{H}} \boldsymbol{H}_{\pi(j)} \boldsymbol{T}_j \boldsymbol{v}_j &= \\ \boldsymbol{u}_i^{\mathrm{H}} \boldsymbol{u}_j \lambda^j_{\bar{k},\bar{s}} &= 0, \end{aligned}$$

which is, as it has been shown, equal to zero due to orthogonality of the receive beamforming vectors. Effective transmission of information occurs over each of the allocated scalar subchannels whose gain is given by

$$g_j = \boldsymbol{u}_j^{\mathrm{H}} \boldsymbol{H}_{\pi(j)} \boldsymbol{v}_j, \quad \forall j.$$

Over this set of virtually decoupled channels allocation of transmit power can be chosen in order to optimize performance.

If applied to a single-user MIMO channel, this algorithm performs a singular value decomposition of the channel matrix independently of $\mathcal{R}$ and, therefore, preserves capacity. If applied to a broadcast channel with single-antenna receivers, this algorithm is equivalent to the zero-forcing dirty-paper coding algorithm discussed in [23, 97, 123]. If applied to a general broadcast channel with multiple antennas at the receivers and each user systematically gets allocated as many dimensions as it can support in consecutive steps, we obtain the block-wise decomposition approach proposed in [108, 41]. Finally, if we restrict the number of dimensions assigned to each user to one, the algorithm proposed in [79] results.

4.1.3.1 Successive subchannel allocation method for linear approaches

The beamforming vectors delivered by SESAM do not suppress interference completely. Rather, the impact of the remaining interference must be eliminated by resorting to a dirty-paper coding scheme, whose efficient implementation is still subject of ongoing research [46, 149, 145, 151, 124]. If, due to practical reasons, independent coding of information streams is preferred, SESAM can still be used as a means of finding a good solution to the general subchannel allocation problem of linear decomposition approaches. To be more specific, SESAM can be used in order to obtain a convenient set of initial receive filter matrices $\tilde{\boldsymbol{U}}_k$, $k = 1, \ldots, K$, based on which linear decomposition can be performed in a subsequent step. Thus, in this case, SESAM is not used to determine the beamforming vectors in an explicit way. Rather, SESAM provides a means to determine which users are served and over how many spatial dimensions, and to conveniently pre-condition the choice of beamforming vectors. The resulting algorithm is given in Algorithm 4.2. In line 7 the new computed receive beamforming vector is incorporated as an additional column to the receive filter matrix of the corresponding user. Different to Algorithm 4.1, where allocation of a new dimension does not have any impact on the previously allocated subchannels, now, a new dimension represents an additional constraint that previously allocated subchannels must fulfil. That is, allocation of a new dimension might be detrimental in terms of performance. Therefore, after assigning a new dimension performance must be evaluated and compared to that obtained after allocation of the previous dimensions.

If performance increases, allocation of a new dimension is considered. If performance decreases, the just computed receive beamforming vector is removed from the corresponding receive filter matrix and no further allocation steps are carried out. Once the allocation process is completed, the broadcast channel is linearly decomposed by first computing matrices $\bar{\boldsymbol{H}}_k$ for every user with the resulting matrices $\tilde{\boldsymbol{U}}_k$ (cf. Eq. 4.3). Then the projectors $\boldsymbol{T}_k^{\perp} = \boldsymbol{I}_t - \bar{\boldsymbol{H}}_k^{\mathrm{H}}(\bar{\boldsymbol{H}}_k\bar{\boldsymbol{H}}_k^{\mathrm{H}})^{-1}\bar{\boldsymbol{H}}_k$ associated to the null space of each of these matrices are computed and, finally, singular value decompositions of the products $\tilde{\boldsymbol{U}}_k^{\mathrm{H}}\boldsymbol{H}_k\boldsymbol{T}_k^{\perp}$ are performed (cf. Section 4.1.1). This algorithm allows a joint treatment of the different aspects of the subchannel allocation problem in the context of linear decomposition approaches such as user selection, number of dimensions assigned to each user and determination of beamforming vectors. The choice of these parameters can be influenced by the performance measure of interest through the selection rule $\mathcal{R}$. Of course, now, the link between the selection rule and the final performance is weaker than in Algorithm 4.1. This is due to the fact that the singular values computed in line 5 do not correspond any more to actual channel gains of potentially allocated subchannels. However, in general, the singular values will still be good estimates of the final channel gains. This is somehow ensured by the performance evaluation carried out within the repeat loop, which prevents from too tough constraints being imposed on already allocated spatial dimensions.

Algorithm 4.2 SESAM-based subchannel allocation for linear decomposition approaches

1: $j \leftarrow 1, \quad \boldsymbol{T}_1 \leftarrow \boldsymbol{I}_t$
2: $\tilde{\boldsymbol{U}}_k = [], \quad k = 1, \ldots, K,$ Receive filter matrices are initialized as empty matrices
3: **repeat**
4: $\quad \boldsymbol{H}_k^j \leftarrow \boldsymbol{H}_k\boldsymbol{T}_j, \quad k = 1, \ldots, K$
5: $\quad \boldsymbol{H}_k^j \leftarrow \boldsymbol{U}_k^j\boldsymbol{\Lambda}_k^j\boldsymbol{V}_k^{j,\mathrm{H}}, \quad k = 1, \ldots, K$
6: $\quad (\bar{k}, \bar{s}) \leftarrow \mathcal{R}\left(\left\{\lambda_{k,s}^j | k = 1, \ldots, K,\ s = 1, \ldots, \rho_k^j\right\}\right),\ \boldsymbol{v}_j \leftarrow \boldsymbol{V}_{\bar{k}}^j\boldsymbol{e}_{\bar{s}},\ \boldsymbol{u}_j \leftarrow \boldsymbol{U}_{\bar{k}}^j\boldsymbol{e}_{\bar{s}}$
7: $\quad \tilde{\boldsymbol{U}}_{\bar{k}} = [\tilde{\boldsymbol{U}}_{\bar{k}} | \boldsymbol{u}_j]$
8: $\quad$ Perform linear decomposition based on $\tilde{\boldsymbol{U}}_k, \quad k = 1, \ldots, K$ (cf. Section 4.1.1)
9: $\quad$ Evaluate performance
10: $\quad$ **if** performance decreases **then**
11: $\quad\quad$ Remove $\boldsymbol{u}_j$ from $\tilde{\boldsymbol{U}}_{\bar{k}}$
12: $\quad\quad$ Break repeat loop
13: $\quad$ **end if**
14: $\quad \boldsymbol{T}_{j+1} \leftarrow \boldsymbol{T}_j - \boldsymbol{v}_j\boldsymbol{v}_j^{\mathrm{H}}, \quad j \leftarrow j + 1$
15: **until** $j > \sum_k r_k$ or $\boldsymbol{T}_j = \boldsymbol{0}$
16: Perform linear decomposition based on $\tilde{\boldsymbol{U}}_k, \quad k = 1, \ldots, K$

In the next sections, adequate selection rules are proposed for the three optimization problems discussed in Chapter 3. Discussion will focus on Algorithm 4.1. However, numerical results will also include performance curves of Algorithm 4.2. In each case the selection rule applied will be the same as that used for the SESAM algorithm.

4.1.4 Sum-rate maximization

4.1.4.1 Selection rule

Aiming at the maximization of the sum-rate, it seems convenient to perform allocation so that at each step the spatial subchannel is selected that being orthogonal to all previously established subchannels exhibits the largest channel gain. Mathematically, the selection rule can be defined as

$$\mathcal{R}\left(\left\{\lambda_{k,s}^{j}|k=1,\ldots,K,\quad s=1,\ldots,\rho_k^j\right\}\right)= \\ =\arg\max_{k,s}\left\{\lambda_{k,s}^{j}|k=1,\ldots,K,\ s=1,\ldots,\rho_k^j\right\}. \tag{4.8}$$

The following theorem provides some rationale for this ordering.

Theorem 4.1.1. *Let g_i, $i=1,\ldots,j$ be the channel gains corresponding to the j first allocated subchannels and let C_j be the sum-rate achieved over these subchannels. Selecting the next subchannel according to the rule proposed in Eq. 4.8 yields the maximum rate increment $\Delta C = C_{j+1} - C_j$.*

Proof. Let $\bar{g}_{j+1} = \max\left\{\lambda_{k,s}^{j+1}|k=1,\ldots,K,\ s=1,\ldots,\rho_k^j\right\}$ be the channel gain resulting from application of the proposed rule in the $j+1$ allocation step and let g_{j+1} be the subchannel gain resulting from the application of any other rule. Let $\bar{\mathcal{G}} = \{g_i|i=1,\ldots,j\}\cup\{\bar{g}_{j+1}\}$ and $\mathcal{G} = \{g_i|i=1,\ldots,j\}\cup\{g_{j+1}\}$

Assume that the optimum waterfilling allocation for the set $\mathcal{G}$ yields the waterfilling level η (cf. Section 4.1.4.2). For the set $\bar{\mathcal{G}}$, consider a suboptimum power allocation where subchannels $i \leq j$ are waterfilled to the level η and subchannel $j+1$ receives the same power as subchannel $j+1$ in $\mathcal{G}$. Obviously, the first j subchannels achieve the same transmission rate in both sets. Transmission rate achieved over subchannel $j+1$ in $\bar{\mathcal{G}}$ will be larger than or equal to that achieved over subchannel $j+1$ in $\mathcal{G}$ as $\bar{g}_{j+1} \geq g_{j+1}$. Waterfilling allocation of power over the subchannels of set $\bar{\mathcal{G}}$ can only lead to an even larger transmission rate for this set. □

Note that $\bar{g}_{j+1}$ in the proof above can be characterized as

$$\bar{g}_{j+1}^2 = \max_k \max_{\boldsymbol{v}} \boldsymbol{v}^{\mathrm{H}}\boldsymbol{H}_k^{j+1,\mathrm{H}}\boldsymbol{H}_k^{j+1}\boldsymbol{v},$$

subject to $\|\boldsymbol{v}\|_1 = 1$. That is, limiting the canditate subchannels to be eigenmodes of the projected channel matrices, as SESAM does, is without loss of optimality in terms of the rate increment that can be attained in one allocation step.

If only single-antenna receivers are considered, SESAM with the allocation rule given in Eq. 4.8 coincides with the algorithm presented in [123]. If for the general setting the number of allocated dimensions per user is restricted to one, the SESAM algorithm with this selection rule coincides with the algorithm proposed in [79]. Of course, the analytical results presented in these works concerning the asymptotic performance behavior at high SNR and for a large number of users immediately apply to SESAM.

In a multicarrier setting, the SESAM algorithm with the selection rule given in Eq. 4.8 can be independently executed on each subcarrier without incurring performance loss

with respect to a direct application of the algorithm using the compact representation of multicarrier channels as block diagonal channel matrices (see Eq. 3.14). This allows for a parallel implementation of the algorithm across subcarriers.

4.1.4.2 Power allocation policy

Let g_j, $j = 1, \ldots, J$ be the channel gains of the subchannels resulting from application of SESAM to a Gaussian broadcast channel. The optimum power allocation policy in terms of sum rate is obtained by solving

$$\arg \max_{P_{1,\ldots,J}} \sum_{j=1}^{J} \log_2(1 + g_j^2 P_j), \tag{4.9}$$

subject to $P_1 + \ldots + P_J \leq P$ and $P_j \geq 0, \forall j$. The resulting optimum policy is the well-known waterfilling power allocation [40], which reads

$$P_j = \max\left\{\eta - \frac{1}{g_j^2}, 0\right\},$$

where η is the waterfilling level, which is chosen to fulfil the transmit power constraint with equality.

4.1.4.3 Numerical results

In this section simulation results are shown corresponding to a multicarrier transmission system with $N = 16$ uncorrelated subcarriers. The basic difference between this setting and a system with $N = 1$ is basically the additional degrees of freedom that the multicarrier system offers for allocation of power over the spectral components. Average performance of both systems would be identical if a uniform allocation of power were applied across subcarriers.

Fig. 4.1 shows average sum capacity curves for a Rayleigh distributed channel with $t = 4$ transmit antennas, $K = 2$ users and $r_1 = r_2 = 2$ antennas at each receiver. The entries in the channel matrix corresponding to a particular user on a particular subcarrier have been assumed to be mutually independent with variance equal to one. The horizontal axis represents the ratio between transmit power per subcarrier and the noise variance at any receive antenna in decibels. The performance gap between SESAM and the optimum approach is for all practical purposes inexistent. In this case, using the fact that the composite broadcast channel matrix $\boldsymbol{H}$ will in general have full row rank, convergence of SESAM and the optimum approach at high SNR can be analytically shown along the lines of [23, Theorem 4] and [41, Theorem 1][4]. At low SNR values, SESAM just serves one of the users

[4]The proof relies on the fact that the product of the eigenvalues of $\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}}$ is equal to the square of the product of the diagonal entries of $\boldsymbol{U}^{\mathrm{H}}\boldsymbol{H}\boldsymbol{V}$, where $\boldsymbol{V}$ is a matrix whose columns are an orthonormal base of the space spanned by the columns of $\boldsymbol{H}^{\mathrm{H}}$, $\boldsymbol{U}$ is a unitary matrix and $\boldsymbol{U}^{\mathrm{H}}\boldsymbol{H}\boldsymbol{V}$ has a triangular structure. The result follows by noting that the transmit and receive weighting vectors computed by SESAM yield such matrices and, at high SNR, capacity is essentially determined by the product of eigenvalues of $\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}}$ in the single user link and the square of the product of the diagonal entries of $\boldsymbol{U}^{\mathrm{H}}\boldsymbol{H}\boldsymbol{V}$ in the broadcast channel.

over the strongest dimension available in the system. The linear decomposition technique behaves exactly as SESAM in the low SNR regime, just transmitting information over the strongest subchannel. Therefore, performance of both suboptimum approaches coincide. Although difficult to show analytically, apparently, the optimum solution adopts the same transmission strategy at low SNR. The performance gap between the linear decomposition approach and the other two curves at high SNR amounts to approximately 2.5 bits. Note that in this case the shaping loss incurred if the popular Tomlinson-Harashima precoding is applied in order to mitigate known interference amounts to 2.03 bits.[5] That is, implementation of successive encoding based on this simple and practical technique for canceling known interference would not provide any significant gain with respect to a plain linear scheme.

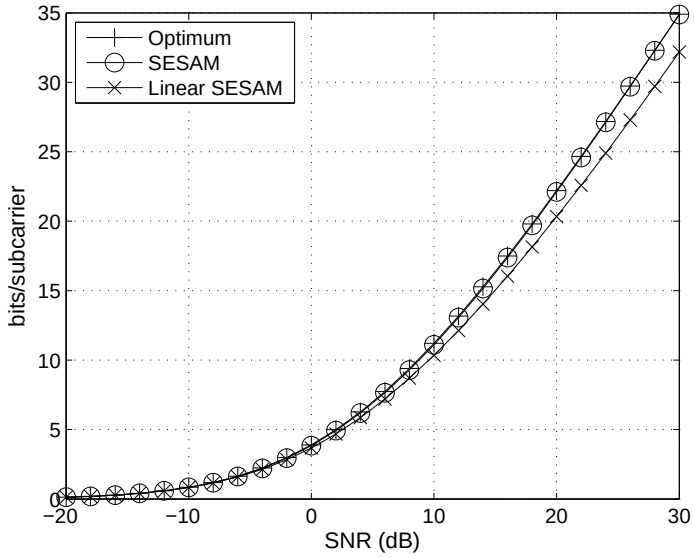

Figure 4.1: Average sum rate for a Gaussian broadcast channel with spatially uncorrelated Rayleigh-fading channel coefficients. $t = 4$, $r_k = 2$, $N = 16$, $K = 2$.

In Fig. 4.2 average sum capacity curves are shown for a scenario as described by the settings used in Fig. 4.1 but where spatial correlation has been introduced on the transmit side. A transmit correlation matrix $\boldsymbol{R}_{\mathrm{Tx}} = \mathrm{E}\{\boldsymbol{H}^{\mathrm{H}}\boldsymbol{H}\}$ has been considered with the following eigenvalue profile,

$$\boldsymbol{\Lambda} = \mathrm{diag}[\ 15,\quad 1,\quad 0,\quad 0\ ]. \tag{4.10}$$

The practical case of two users being in locations few meters apart from each other that are reached by the base station through quite a narrow bundle of angles of departure matches the setting proposed here. The asymptotic slope of all curves is just half of that

[5]The shaping loss per real dimension amounts to 0.254 bits [46]. This number must be multiplied by 2 in order to get the shaping loss per complex dimension. The resulting number must be multiplied by the number of dimensions allocated in order to get the total loss per channel use.

corresponding to the curves displayed in Fig. 4.1. This is due to the reduced rank of the channel, which is now 2 rather than 4. As before, almost no gap can be seen between SESAM and the optimum approach. Note, however, that now the composite channel matrix is not full row rank. Therefore, analyses based on this assumption as those carried out in [23, 41] are not applicable. The performance gap between the successive encoding schemes and the linear decomposition approach can be only observed at high SNR values and, even there, it is rather small.

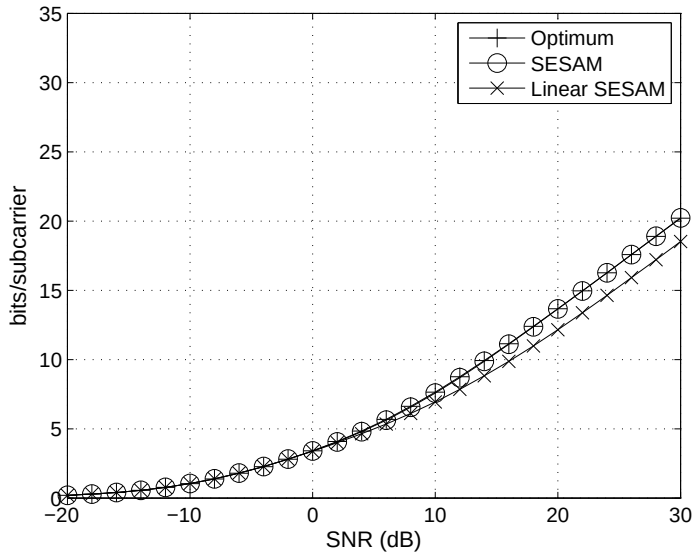

Figure 4.2: Average sum rate for a Gaussian broadcast channel with spatially correlated Rayleigh-fading channel coefficients. $t = 4$, $r_k = 2$, $N = 16$, $K = 2$.

Fig. 4.3 shows average sum capacity curves for a Rayleigh distributed channel with $t = 4$ transmit antennas, $K = 10$ users and $r_k = 2$ antennas at each receiver. Entries in the composite channel matrix of each subcarrier are assumed to be mutually independent and with covariance equal to one. Different from the settings of Figs. 4.1 and 4.2, now, the total number of receive antennas in the system is larger than the number of transmit antennas. This calls for a decision regarding the users to be served and the number of subchannels to be assigned to these users on a particular subcarrier. This additional degree of freedom, commonly known as multiuser diversity, is exploited by all the simulated schemes and results in improved performance as compared to previous figures (cf. [79]). The gap between the linear decomposition approach and successive encoding approaches becomes smaller. That is, linear schemes benefit from multiuser-diversity more than successive-encoding-based decomposition approaches. This is due to the fact that as the number of users increases, it is easier to find a set of t quasi-orthogonal spatial dimensions to allocate.

Fig. 4.4 shows average sum capacity curves for a scenario as described by the settings used in Fig. 4.3 but where correlation has been introduced between transmit antenna elements. For these simulations, an eigenvalue profile of the transmit covariance matrix

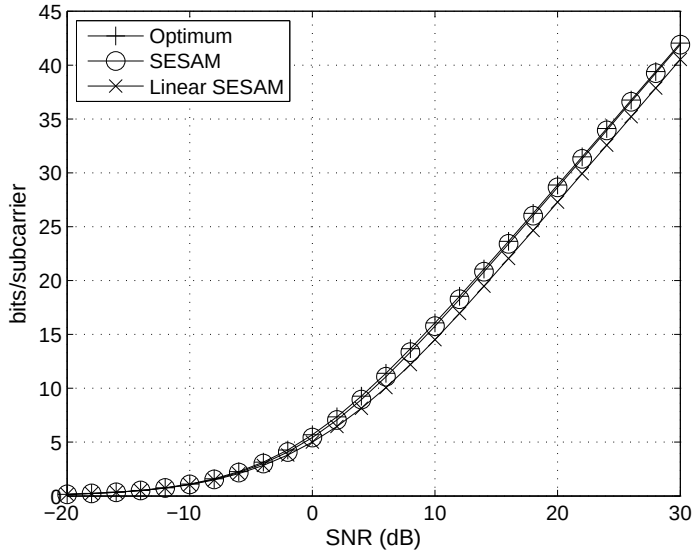

Figure 4.3: Average sum rate for a Gaussian broadcast channel with spatially uncorrelated Rayleigh-fading channel coefficients. $t = 4$, $r_k = 2$, $N = 16$, $K = 10$.

has been considered proportional to

$$\boldsymbol{\Lambda} = \text{diag}[\ 10,\ \ 5,\ \ 1,\ \ 0\ ]. \tag{4.11}$$

This profile may very well match a scenario in which a group of users located in a same certain area, such as a square or street, are reached from the base station over the same moderately broad bundle of angles of departure. Again, the rank deficiency of the channel causes a decay of the asymptotic growth of all approaches. As in previous figures SESAM shows an insignificant performance loss with respect to the optimum solution. Also a mild gap can be appreciated between the optimum approach and the linear scheme at moderate and high SNR values.

SNR (dB)	-14	-10	-6	-2	2	6	10	14	18	22	26
Uncorrelated, K = 2	1.2	1.1	1.1	1.1	1.4	1.7	1.6	1.4	1.2	1.0	1.0
Uncorrelated, K = 10	1.1	1.1	1.1	1.3	2.0	2.7	3.9	5.2	6.1	6.3	6.4
Correlated, K = 2	46.0	51.0	55.3	12.5	3.8	3.0	3.6	5.1	6.8	8.1	9.0
Correlated, K = 10	3.3	2.6	2.7	3.2	3.0	3.5	4.5	5.7	6.6	7.3	7.6

Table 4.1: Average number of iterations needed by Algorithm 3.6 to achieve $0.999 R_{\text{SESAM}}$.

Table 4.1 shows the average number of iterations required by the optimum iterative Algorithm 3.6 in order to reach 99.9% of the sum rate achieved by SESAM (R_{SESAM}). Numbers range between almost one iteration for uncorrelated scenarios at low SNR and more than 55 iterations at -6 dB, $K = 2$ and correlated channels. This indicates that the additional computational complexity of optimal iterative approaches relative to SESAM strongly depends on the particular setting.

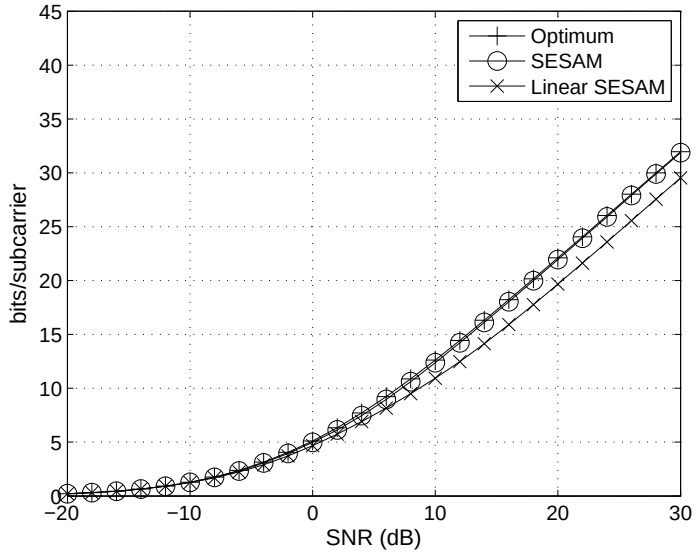

Figure 4.4: Average sum rate for a Gaussian broadcast channel with spatially correlated Rayleigh-fading channel coefficients. $t = 4$, $r_k = 2$, $N = 16$, $K = 10$.

4.1.5 Weighted sum-rate maximization

4.1.5.1 Selection rule

Given a set of weights $\mu_k \in \mathbb{R}_+$, $k = 1, \ldots, K$, these can easily be incorporated into the selection rule in order to perform a priority-sensitive subchannel allocation as follows,

$$\mathcal{R}\left(\left\{\lambda_{k,s}^j | k = 1, \ldots, K, \quad s = 1, \ldots, \rho_k^j\right\}\right) = \\ = \arg\max_{k,s} \left\{\mu_k \lambda_{k,s}^j | k = 1, \ldots, K, \ s = 1, \ldots, \rho_k^j\right\}. \tag{4.12}$$

If all weights are equal this rule is identical to the sum-rate maximizing rule in Eq. 4.8. Setting all but the weight of a particular user equal to zero, this rule allocates all dimensions to the only user whose weight is different from zero. In this case the SESAM algorithm performs a SVD of the single user channel, which is capacity preserving. In all intermediate cases, this rule increases the probability of those users getting dimensions allocated that have a higher priority. This is done by performing a kind of trade-off between priorities on the one hand and channel strength on the other. In a multicarrier setting, the SESAM algorithm with the selection rule given in Eq. 4.12 can be independently executed on each subcarrier without incurring performance loss with respect to a direct application of the algorithm using the compact representation of multicarrier channels as block diagonal channel matrices (see Eq. 3.14). This allows for a parallel implementation of the algorithm across subcarriers.

4.1.5.2 Power allocation policy

Let g_j, $j = 1, \ldots, J$ be the channel gains of the subchannels resulting from application of SESAM to a Gaussian broadcast channel. The optimum power allocation policy in terms of weighted sum-rate is obtained by solving

$$\arg \max_{P_{1,\ldots,J}} \sum_{j=1}^{J} \mu_{\pi(j)} \log(1 + g_j^2 P_j), \tag{4.13}$$

subject to $P_1 + \ldots + P_J \leq P$ and $P_j \geq 0, \forall j$. Choosing all priorities equal, i.e. $\mu_1 = \ldots = \mu_K$, Problem 4.13 reduces to Problem 4.9, for which the solution is the well-known waterfilling power allocation. In the general case, the solution can be derived by solving the KKT conditions, which are in this case sufficient due to the concavity of the objective function and the convexity of the feasible set. The Lagrangian of Problem 4.13 is given by

$$L(P_1, \ldots, P_J, \nu_0, \ldots, \nu_J) = \sum_{j=1}^{J} \mu_{\pi(j)} \log(1 + g_j^2 P_j) + \sum_{j=1}^{J} \nu_j P_j + \nu_0 \left(P - \sum_{j=1}^{J} P_j \right).$$

The corresponding KKT conditions can be written as follows,

$$\frac{\partial L}{\partial P_j} = \frac{\mu_{\pi(j)} g_j^2}{1 + P_j g_j^2} - \nu_0 + \nu_j = 0, \quad j = 1, \ldots, J, \tag{4.14}$$

$$\sum_{j=1}^{J} P_j \leq P, \quad \nu_0 \geq 0, \quad P_j \geq 0, \ \nu_j \geq 0, \ j = 1, \ldots, J,$$

$$\nu_0 \left(P - \sum_{j=1}^{J} P_j \right) = 0, \quad \nu_j P_j = 0, \ j = 1, \ldots, J. \tag{4.15}$$

From Eq. 4.14 follows

$$P_j = \frac{\mu_{\pi(j)}}{\nu_0 - \nu_j} - \frac{1}{g_j^2}, \ j = 1, \ldots, J. \tag{4.16}$$

If $P > 0$, there will optimally be at least a subchannel j that gets some power allocated, i.e., $P_j > 0$ for at least one subchannel. Considering Eq. 4.16, this necessarily implies that $\nu_0 > 0$, from which, due to the first slackness condition in Eqs. 4.15, equality in the total power constraint follows, i.e.,

$$\sum_{j=1}^{q} P_j = P. \tag{4.17}$$

Let $\mathcal{S}$ be a set of indices j such that $P_j > 0$ if $j \in \mathcal{S}$ and $P_j = 0$ if $j \notin \mathcal{S}$. From the second slackness condition in Eqs. 4.15, it follows that $\nu_j = 0$ for $j \in \mathcal{S}$. Now, defining the waterlevel $\eta = 1/\nu_0$, and using Eq. 4.16 the optimum power allocation can finally be written as

$$P_j = \max \left\{ \eta \mu_j - \frac{1}{g_j^2}, 0 \right\}, \ j = 1, \ldots, J,$$

where η must be chosen to satisfy Eq. 4.17. This solution can be viewed as a kind of generalized waterfilling [90].

4.1.5.3 Numerical results

In this section numerical results are presented that correspond to a multicarrier transmission system with $N = 16$ uncorrelated subcarriers. Again, the four different settings of Section 4.1.4.3 are considered, which differ in the number of users and the spatial correlation properties. For given weights, we plot the average rates obtained by the users in the network, where averaging takes place over a number of different channel realizations. In order to compute optimum rate vectors, Algorithm 3.10 has been applied in conjunction with Algorithm 3.9. In order to save simulation time the number of iterations of Algorithm 3.10 has been limited to 100. For Algorithm 3.9 the stop criterion parameter ϵ has been chosen to be 10^{-5}.

In Fig. 4.5 average rate pairs are shown for a Rayleigh distributed broadcast channel with $t = 4$ transmit antennas, $K = 2$ users and $r_1 = r_2 = 2$ antennas at each receiver. The entries in the channel matrix corresponding to a particular user on a particular subcarrier have been assumed to be mutually independent with variance equal to one. Average rate pairs have been plotted for three different SNR values and weight pairs (μ_1, μ_2) with $\mu_2 = 1-\mu_1$, $\mu_1 = n/10$ and $n \in \{1, 2, \ldots, 9\}$. Both the rate pairs corresponding to the optimum solution and the rate pairs achieved by SESAM seem to lie on the same curved line that may be viewed as the boundary of the averaged capacity region. That is, SESAM is able not only of practically achieving sum capacity but also of closely approximating any other points on the boundary of the capacity region. In terms of weighted sum rate, the maximum gap between SESAM and the solution achieved by the optimum algorithm amounts to 1.44% of the optimum value at 15 dB for the weight pairs $(0.4, 0.6)$ and $(0.6, 0.4)$. Despite this insignificant difference in terms of weighted sum-rate, for a given weight pair, the distribution of rates among users remarkably differs, being the points attained by SESAM more evenly spaced over the boundary of the region than those delivered by the optimum algorithm. The gap between the linear SESAM scheme and the other two approaches is especially visible at 15 and 25 dB. The points resulting from this scheme describe a curve that is strictly in the interior of the region delimited by the points of the other two approaches. However, the gap in terms of weighted sum-rate is surprisingly small reaching a maximum of 9% with respect to the optimum at 25 dB for the weight pairs $(0.4, 0.6)$ and $(0.6, 0.4)$.

Table 4.2 shows the average number of iterations required to reach 99.9% of the weighted sum rate attained by SESAM. Outer iterations refer to the number of iterations performed by Algorithm 3.10. Inner iterations refer to the average number of iterations performed by Algorithm 3.9 per subcarrier in each iteration of Algorithm 3.10. As mentioned before, the maximum number of outer iterations has been limited to 100, therefore, some of these numbers would be much larger if this limitation had not been imposed. In fact, due to this limitation in the number of iterations, in some of the settings the 99.9% of the SESAM performance is not reached by the iterative algorithms. For instance, the weighted sum rate achieved by the iterative algorithm at 5 dB for $\mu_1 = 0.5$ is just 98.3% of the value achieved by SESAM. All in all, the values displayed in Table 4.2 clearly confirm that the number of iterations, and, therefore, the computational complexity, required to compute the optimum solution is extremely dependent on the particular system parameters in a way that is virtually impossible to discern beforehand.

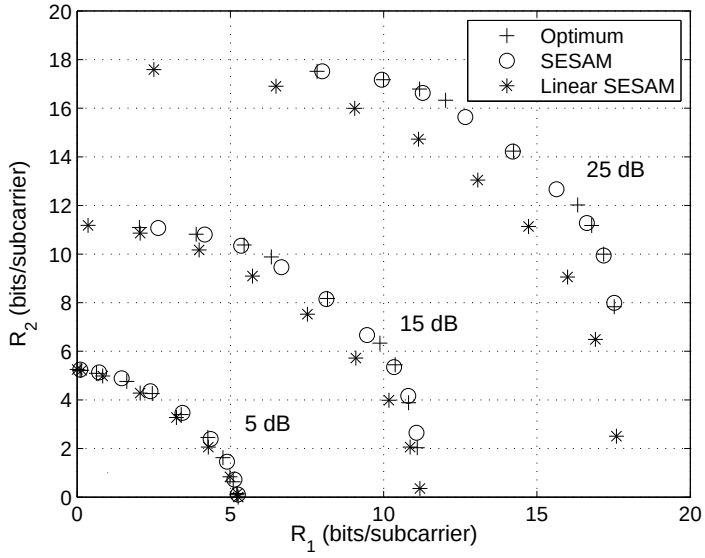

Figure 4.5: Average rate tuples for a Gaussian broadcast channel with spatially uncorrelated Rayleigh-fading channel coefficients. $t = 4$, $r_k = 2$, $N = 16$, $K = 2$.

μ_1	0.1	0.2	0.3	0.4	0.5
inner iterations	1.1/2.0/7.5	1.2/4.1/6.1	2.3/11.7/5.4	3.2/10.8/5.6	1.4/6.0/4.3
outer iterations	98.0/96.2/45.9	98.6/82.0/5.3	89.9/3.6/1.0	80.5/1.0/1.0	95.7/6.6/1.8

Table 4.2: Average numbers of iterations involved in the computation of the optimum solution in order to reach 99.9% of the weighted sum rate achieved by SESAM in a spatially uncorrelated broadcast channel with $t = 4$, $r_k = 2$, $N = 16$, $K = 2$ and SNR = 5/15/25 dB.

In Fig. 4.6 average rate pairs are shown for a Rayleigh distributed broadcast channel with $t = 4$ transmit antennas, $K = 2$ users and $r_1 = r_2 = 2$ antennas at each receiver. The entries in the channel matrix corresponding to a particular user on a particular subcarrier are still assumed to be mutually independent with variance equal to one but channel coefficients corresponding to different transmit antennas are now correlated. The transmit covariance matrix exhibits an eigenvalue profile as given by Eq. 4.10. The SNR values and the weight pairs for which average rate pairs have been computed are the same as in Fig. 4.5. Due to the effect of correlation, all rates are in this case lower than in the previous figure. Due to the limitation in the number of iterations to 100, the iterative algorithm yields a solution that is visibly outperformed by SESAM for $\mu_1 = 0.5$. To be precise, for these weights the iterative algorithm only reaches 90.1% of the weighted sum rate achieved by SESAM at 5 dB. At 15 and 25 dB these numbers are 94.8% and 96.7%, respectively. The maximum performance gap in terms of weighted sum rate between the iterative algorithm and SESAM is observed at 25 dB for the weight pairs $(0.3, 0.7)$ and $(0.7, 0.3)$. For these parameters, SESAM performs 2.53% below the weighted sum rate

achieved by the iterative algorithm. The maximum performance gap between the iterative algorithm and the linear SESAM scheme is observed at 25 dB for the weight pairs $(0.4, 0.6)$ and $(0.6, 0.4)$ and amounts to 10.47% of the weighted sum rate achieved by the optimum algorithm. Particularly remarkable is the even distribution of the points corresponding to SESAM and the linear scheme over the boundary of the respective regions. This is in sharp contrast to the high concentration of optimum points close to the axes.

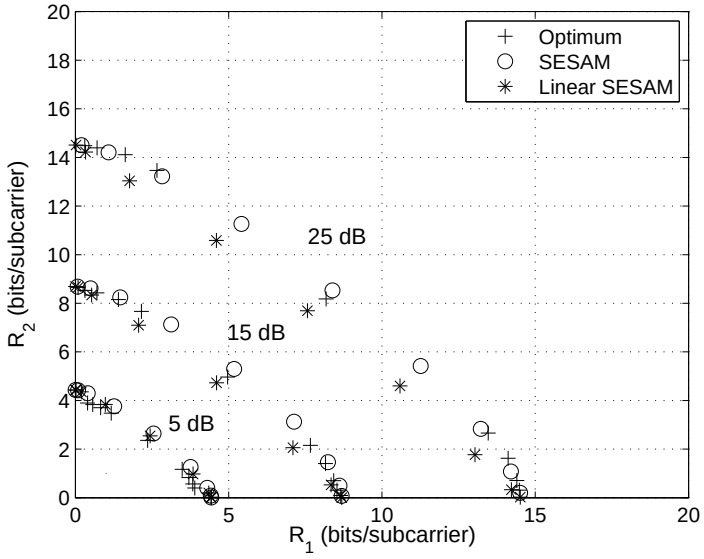

Figure 4.6: Average rate tuples for a Gaussian broadcast channel with spatially correlated Rayleigh-fading channel coefficients. $t = 4$, $r_k = 2$, $N = 16$, $K = 2$.

For completeness we include Table 4.3, which shows the average number of iterations required to reach 99.9% of the weighted sum rate attained by SESAM. As mentioned before, due to the limitation in the number of iterations, performance of the iterative approach does actually not reach this value in some cases. In these cases, the actual number of iterations required to reach it would be significantly larger. As we already observed in Table 4.1 correlations apparently lead to an increase in the number of iterations required by optimum iterative approaches in order to reach convergence.

μ_1	0.1	0.2	0.3	0.4	0.5
inner iterations	1.0/1.2/6.7	1.1/1.7/9.5	1.1/4.2/24.8	1.1/4.5/22.7	1.1/1.2/1.7
outer iterations	99.7/95.8/35.0	99.7/93.1/14.0	99.8/78.8/1.8	100/74.2/3.1	100/100/100

Table 4.3: Average numbers of iterations involved in the computation of the optimum solution in order to reach 99.9% of the weighted sum rate achieved by SESAM in a spatially correlated broadcast channel with $t = 4$, $r_k = 2$, $N = 16$, $K = 2$ and SNR = 5/15/25 dB.

Fig. 4.7 shows average rates obtained by the $K = 10$ users of a Rayleigh distributed broadcast channel with $t = 4$ transmit antennas and $r_k = 2$ antennas at each receiver

when weighted sum-rate maximization is performed with weights $\mu_k = k$, $k = 1, \ldots, K$ at SNR = 15 dB. The entries in the channel matrices are assumed to be uncorrelated and with variance 1.The weighted sum rate achieved by SESAM is solely 2.1% below the value achieved by the optimum algorithm. In turn, the gap between the linear approach and the optimum algorithm amounts to just 10.2% of the optimum weighted sum rate. In order to reach 99.9% of the weighted sum rate achieved by SESAM the optimum iterative algorithm requires 1.32 outer iterations in average and 12.54 inner iterations per subcarrier and outer iteration. Again we observe that even if the difference between SESAM and the optimum solution is very small in terms of weighted sum rate, the resulting distributions of rates among users are quite different. In Fig. 4.8 the same broadcast channel is considered as in Fig. 4.7 but with channel coefficients corresponding to different transmit antennas being correlated. The transmit correlation matrix has an eigenvalue profile proportional to that given by Eq. 4.11. The weights are as defined above and average rates achieved by each user are represented for a SNR = 15 dB. Due to correlations the rates are lower than in Fig. 4.7. SESAM achieves a weighted sum rate that is just 1.43% below that delivered by the optimum iterative algorithm. The performance loss of the linear decomposition scheme with respect to the optimum algorithm amounts to 12.87%. In order to reach 99.9% of the weighted sum rate achieved by SESAM the optimum iterative algorithm requires 10.87 outer iterations in average and 11.92 inner iterations per subcarrier and outer iteration, which again confirms the fact that correlations cause an increase in the number of iterations required to reach convergence.

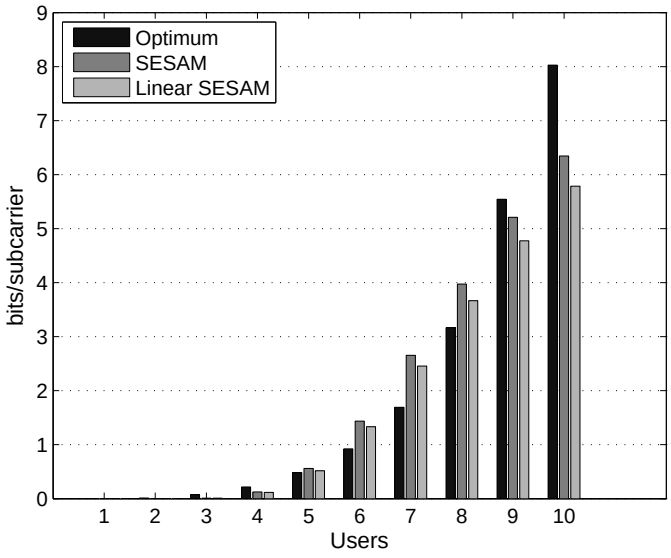

Figure 4.7: Average rate tuple for a Gaussian broadcast channel with spatially uncorrelated Rayleigh-fading channel coefficients. $t = 4$, $r_k = 2$, $N = 16$, $K = 10$, SNR = 15 dB.

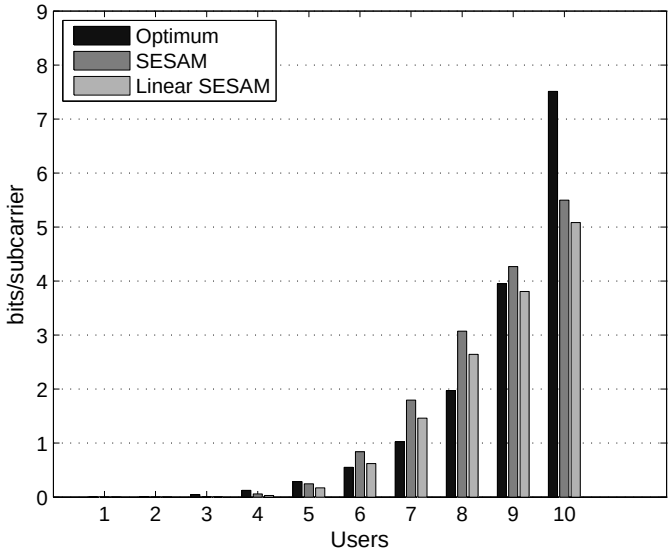

Figure 4.8: Average rate tuple for a Gaussian broadcast channel with spatially correlated Rayleigh-fading channel coefficients. $t = 4$, $r_k = 2$, $N = 16$, $K = 10$, SNR = 15 dB.

4.1.6 Rate balancing

There is an important qualitative difference between the rate balancing problem and the sum-rate or weighted sum-rate problems discussed in previous sections. While the latter do not impose any constraint with respect to the number of users that get served in the network, the former requires that all users whose relative rate requirements q_k (cf. Section 3.3) are larger than zero be necessarily served. Obviously, if decomposition approaches are considered, this is only possible provided that the total amount of dimensions in the system exceeds the number of users with positive rate requirements in the network. To ensure this, beyond spatial dimensions additional time and frequency dimensions might have to be considered. In the following, we consider a MIMO OFDM broadcast channel and we assume that the number of subcarriers is larger than the number of users with positive rate requirements in the network.

4.1.6.1 Selection rule

As in the treatment of the sum-rate and weighted sum-rate problems, SESAM is applied separately on each subcarrier. However, now, rather than applying a selection rule to each subcarrier independently, allocation on a particular subcarrier takes into account allocation on all other subcarriers. That is, we have a selection rule that coordinates the allocation of new spatial dimensions across subcarriers. This is needed so as to enforce that at each allocation step of the SESAM algorithm users get a share of resources according to their rate requirements. Correspondingly, the selection rule is not any more a simple operator on the set of singular values obtained on each particular subcarrier but a more sophisticated procedure that takes into account all singular values across the spectrum and the relative

rate requirements of the users in the network. Aiming at the allocation of the jth spatial component on every subcarrier, the first and second steps of the SESAM algorithm[6] are independently executed on all frequency dimensions. Let $\boldsymbol{\Lambda}_{n,k}^{j}$ be the matrix of singular values of user k, on subcarrier n, obtained during the jth execution of the repeat loop of SESAM (cf. Algorithms 4.1 and 4.2),[7] and let $\lambda_{n,k,s}^{j}$ be the sth singular value of this matrix. The selection procedure that we propose consists of three basic steps.

First, for each user, the largest singular value on each subcarrier is selected, i.e.,

$$\lambda_{n,k}^{j} = \max_{s}\{\lambda_{n,k,s}^{j}\} \quad \forall n, k,$$

and only these subchannels are considered in the following steps of the allocation rule.

Second, the number of frequency components is determined that shall be assigned to each user taking into account a given vector $\boldsymbol{q}$ of relative rate requirements. To this end, first, the capacity is computed that each user could achieve in this allocation layer should all frequency components be assigned to that user. For example, capacity of user k is computed as

$$C_{k}^{j} = \frac{1}{N}\sum_{n=1}^{N} \log\left(1 + P_{n,k}(\lambda_{n,k}^{j})^{2}\right),$$

where $P_{n,k}$ is obtained from a waterfilling power allocation over the singular values $\lambda_{n,k}^{j}$. In order to compute capacities at layer j, it is assumed that the power is limited to P/j. This is merely a heuristic that permits computation of capacity in a particular layer without considering subchannels assigned in previous or subsequent allocation steps. The reason for the division by j is that channel gains become smaller in each layer and so does the power finally allocated to each layer. Then, we consider the affine space defined by the rate vectors $\boldsymbol{\rho}_k = C_k^j \boldsymbol{e}_k$, $\forall k$, where $\boldsymbol{e}_k$ is a column vector of dimension K with a 1 on its kth row and zeros elsewhere, and compute the intersection point of this space and the straight line defined by the given vector of relative rate requirements $\boldsymbol{q}$. The equation of the affine space is given by $\boldsymbol{\rho} = \sum_{k=1}^{K} \beta_k \boldsymbol{\rho}_k$ with $\sum_{k=1}^{K} \beta_k = 1$, and that of the straight line by $\boldsymbol{\rho} = \gamma \boldsymbol{q}$. The intersection point is obtained solving the following linear system of equations,

$$\begin{aligned} \gamma \boldsymbol{q} &= \beta_1 \boldsymbol{\rho}_1 + \beta_2 \boldsymbol{\rho}_2 + \ldots + \beta_K \boldsymbol{\rho}_K \\ 1 &= \beta_1 + \beta_2 + \ldots + \beta_K. \end{aligned}$$

The resulting weight β_k is seen as the fraction of subcarriers that should be allocated to user k at layer j in order to comply with the QoS constraint represented by $\boldsymbol{q}$. Correspondingly, the number of subcarriers assigned to that user in that layer is given by $N_k = \beta_k N$, which can be rounded and readjusted to obtain a set of integral values adding up to the total number of subcarriers. This procedure and interpretation of the parameters β_k is optimum if there is only one spatial dimension, e.g., $t = 1$, the channels are non-frequency-selective, each subcarrier is exclusively assigned to a unique user and the same amount of power is

[6] These are steps 3 and 4 in Algorithm 4.1 and steps 4 and 5 in Algorithm 4.2.

[7] In the following, the jth execution of the repeat loop of SESAM will be occasionally referred to as layer j.

allocated on every subcarrier. Only in such a case the afore mentioned space represents the boundary of the set of achievable rates and the intersection of this space and the straight line defined by vector $\boldsymbol{q}$ is the optimum operational point. Even though in all other cases this method is suboptimum, it shall be seen that it delivers excellent results.

In the third step, allocation of subcarriers to users is performed such that compliance with the subcarrier numbers obtained in the previous step is guaranteed. To this end, first, on each subcarrier the subchannel is selected with largest gain, i.e.,

$$\lambda_n^j = \max_k \left\{ \lambda_{n,k}^j \right\} \ \forall n. \tag{4.18}$$

This selection is optimum with respect to sum capacity but it might not be in agreement with the numbers of subcarriers computed in the previous step. If this is the case the selection must be modified in order to match these numbers. This can be done as follows. Let $\tilde{N}_k$ be the number of subchannels of user k selected according to Eq. 4.18 and define the following sets:

$$\begin{aligned} \mathcal{R} &= \{k | N_k - \tilde{N}_k > 0\}, \\ \mathcal{D} &= \{k | N_k - \tilde{N}_k < 0\}, \\ \mathcal{C} &= \{n | \lambda_n^j = \lambda_{n,k}^j, \ k \in \mathcal{D}\}. \end{aligned}$$

$\mathcal{R}$ is the set of users to which additional subchannels should be assigned. $\mathcal{D}$ is the set of users from which subchannels should be removed. $\mathcal{C}$ is the set of subcarriers on which a user of set $\mathcal{D}$ has been assigned a subchannel. Additionally, we define a set including the difference between gains of selected subchannels and gains of non-selected subchannels, $\mathcal{S} = \{\Delta\lambda_{n,k} | k \in \mathcal{R}, \ n \in \mathcal{C}\}$, where $\Delta\lambda_{n,k} = \lambda_n^j - \lambda_{n,k}^j$. With these definitions the following procedure is repeated until the sets $\mathcal{D}$ and $\mathcal{R}$ are empty, i.e., until the number of subcarriers assigned to each user coincides with the number N_k previously computed.

1. Find the user of set $\mathcal{R}$ and carrier of set $\mathcal{C}$ corresponding to the smallest gain difference with respect to a selected subchannel,

$$(n', k') = \underset{n,k}{\operatorname{argmin}} \left\{ \Delta\lambda_{n,k} \right\}, \quad \Delta\lambda_{n,k} \in \mathcal{S}.$$

2. Find the user to which initially the subchannel on subcarrier n' has been assigned,

$$k'' = \underset{k}{\operatorname{argmax}} \{ \lambda_{n',k}^j \}, \quad k \in \mathcal{D}.$$

3. Change the assignment on the selected subcarrier, i.e. $\lambda_{n'}^j = \lambda_{n',k'}^j$.

4. Update subchannel counters, $\tilde{N}_{k''} = \tilde{N}_{k''} - 1$, $\tilde{N}_{k'} = \tilde{N}_{k'} + 1$, and redefine sets accordingly.

Though suboptimal, this procedure yields a good performance and has a clear rationale. It departs from the sum capacity optimum subchannel selection and modifies at each step the allocation so that the incurred channel gain loss is minimized.

Once allocation at step j has been completed, projectors are correspondingly updated on each subcarrier (line 6 of Algorithm 4.1) and allocation of the $(j+1)$th spatial dimension starts. A summary of the steps involved in this allocation rule is given in Algorithm 4.3.

Algorithm 4.3 Rate balancing allocation rule at layer j

$\lambda_{n,k}^{j} \leftarrow \max_{s}\{\lambda_{n,k,s}^{j}\}, \quad \forall n,k$

$C_k^j \leftarrow \frac{1}{N}\sum_{n=1}^{N}\log\left(1+P_{n,k}(\lambda_{n,k}^{j})^2\right), \quad \forall k$

$\boldsymbol{\rho}_k \leftarrow C_k^j \boldsymbol{e}_k, \quad \forall k$

Solve $\gamma \boldsymbol{q} = \beta_1\boldsymbol{\rho}_1 + \beta_2\boldsymbol{\rho}_2 + \ldots + \beta_K\boldsymbol{\rho}_K$, $1 = \beta_1 + \beta_2 + \ldots + \beta_K$

$N_k \leftarrow \beta_k N, \quad \forall k$

Adjust N_k, $k = 1, \ldots, K$, to obtain integer numbers adding to N

$\lambda_n^j \leftarrow \max_k \left\{\lambda_{n,k}^j\right\}, \ \forall n$

$\tilde{N}_k \leftarrow$ Number of subcarriers for which $\lambda_n^j = \lambda_{n,k}^j$

Define $\mathcal{R} = \{k | N_k - \tilde{N}_k > 0\}$, $\mathcal{D} = \{k | N_k - \tilde{N}_k < 0\}$, $\mathcal{C} = \{n | \lambda_n^j = \lambda_{n,k}^j, \ k \in \mathcal{D}\}$

Define $\mathcal{S} = \{\Delta\lambda_{n,k} | k \in \mathcal{R}, \ n \in \mathcal{C}\}$ with $\Delta\lambda_{n,k} = \lambda_n^j - \lambda_{n,k}^j$

while $\mathcal{R} \neq \emptyset$ and $\mathcal{D} \neq \emptyset$ **do**

  $(n', k') \leftarrow \operatorname{argmin}_{n,k}\{\Delta\lambda_{n,k}\}, \quad \Delta\lambda_{n,k} \in \mathcal{S}$

  $k'' \leftarrow \operatorname{argmax}_{k}\{\lambda_{n',k}^j\}, \quad k \in \mathcal{D}$

  $\lambda_{n'}^j \leftarrow \lambda_{n',k'}^j$

  $\tilde{N}_{k''} \leftarrow \tilde{N}_{k''} - 1$

  $\tilde{N}_{k'} \leftarrow \tilde{N}_{k'} + 1$

end while

4.1.6.2 Power allocation policy

After the allocation process has been concluded, for each user, a set of scalar mutually decoupled subchannels is obtained over which power loading can be applied so as to maximize sum rate under consideration of the given QoS constraint. A suboptimum algorithm for this problem has been previously proposed in [103]. An optimum algorithm is derived in this section. Let $g_{k,\ell}$ represent the channel gain of the ℓth subchannel assigned to user k and L_k the total number of subchannels assigned to that user. The optimization problem to be solved in order to find the power loading that maximizes sum rate subject to a QoS constraint $\boldsymbol{q}$ can be stated as follows,[8]

$$\max_{\boldsymbol{p}_{k=1,\ldots,K}} \frac{1}{q_1} \sum_{\ell=1}^{L_1} \log(1 + P_{1,\ell} g_{1,\ell}^2),$$

subject to

$$\frac{1}{q_k} \sum_{\ell=1}^{L_k} \log(1 + P_{k,\ell} g_{k,\ell}^2) - \frac{1}{q_1} \sum_{\ell=1}^{L_1} \log(1 + P_{1,\ell} g_{1,\ell}^2) = 0, \ \forall k > 1,$$

$P_{k,\ell} \geq 0, \ \forall k, \ell$ and $P - \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} P_{k,\ell} \geq 0$, where $\boldsymbol{p}_k = [\ P_{k,1} \ \ \ldots \ \ P_{k,L_k}\ ]^{\mathrm{T}}$ and $P_{k,\ell}$ is the power allocated to the ℓth subchannel of user k. The Lagrangian of this optimization problem can be written as

$$\begin{aligned} &L\left(P_{\substack{1,\ldots,K \\ 1,\ldots,L_k}}, \eta, \mu_{\substack{1,\ldots,K \\ 1,\ldots,L_k}}, \nu_{1,\ldots,K}\right) = \\ &= \sum_{k=1}^{K} \frac{\nu_k}{q_k} \sum_{\ell=1}^{L_k} \log(1 + P_{k,\ell} g_{k,\ell}^2) + \eta \left(P - \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} P_{k,\ell}\right) + \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} \mu_{k,\ell} P_{k,\ell}, \end{aligned}$$

where $\nu_1 = 1 - \sum_{k=2}^{K} \nu_k$. The corresponding KKT conditions read

$$\frac{\nu_k}{q_k} \frac{g_{k,\ell}^2}{1 + P_{k,\ell} g_{k,\ell}^2} - \eta + \mu_{k,\ell} = 0, \ \forall k, \tag{4.19}$$

$$\frac{1}{q_k} \sum_{\ell=1}^{L_k} \log(1 + P_{k,\ell} g_{k,\ell}^2) - \frac{1}{q_1} \sum_{\ell=1}^{L_1} \log(1 + P_{1,\ell} g_{1,\ell}^2) = 0, \ \forall k > 1, \tag{4.20}$$

$$P - \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} P_{k,\ell} \geq 0, \quad P_{k,\ell} \geq 0, \ \forall k, \ell, \quad \eta \geq 0, \quad \mu_{k,\ell} \geq 0, \ \forall k, \ell, \tag{4.21}$$

$$\eta \left(P - \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} P_{k,\ell}\right) = 0, \quad \mu_{k,\ell} P_{k,\ell} = 0, \ \forall k, \ell. \tag{4.22}$$

Eq. 4.19 can be rewritten as

$$P_{k,\ell} = \frac{\nu_k}{q_k(\eta - \mu_{k,\ell})} - \frac{1}{g_{k,\ell}^2}, \ \forall k, \ell. \tag{4.23}$$

[8] Without loss of generality the rate of user 1 is taken as a reference.

As soon as $P > 0$, optimality implies that at least some subchannels get some power allocated, i.e., $P_{k,\ell} > 0$ for at least some subchannels. From Eq. 4.23 we observe that this forces $\eta > 0$. From this, due to the first slackness condition in Eqs. 4.22,

$$P - \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} P_{k,\ell} = 0 \tag{4.24}$$

follows. That is, the total power constraint must be satisfied with equality. Now, taking $P_{k,\ell} \geq 0$ and $\mu_{k,\ell} P_{k,\ell} = 0$, $\forall k, \ell$, into account we can write

$$P_{k,\ell} = \max \left\{ \xi_k - \frac{1}{g_{k,\ell}^2}, 0 \right\}, \ \forall k, \ell, \tag{4.25}$$

where $\xi_k = \nu_k / q_k \eta$. This result has the form of a waterfilling solution with a user dependent water level ξ_k. These levels must be determined so that the $K - 1$ equalities in Eqs. 4.20 hold and Eq. 4.24 is satisfied with equality. Fortunately, it turns out that there is a unique set of parameters ξ_k, $k = 1, \ldots, K$ that satisfy these conditions and, therefore, even though the optimization problem is non-convex, the KKT conditions are sufficient. This can be seen as follows. Let $\mathcal{S}_k(\xi_k)$ be a set including all indexes $\ell \in \{1, \ldots, L_k\}$ of subchannels of user k such that $P_{k,\ell} > 0$. Fixing ξ_1, considering Eq. 4.25 and writing the sum of logarithms as the logarithm of a product in Eqs. 4.20 we obtain

$$\prod_{\ell \in \mathcal{S}_k(\xi_k)} \left(\xi_k g_{k,\ell}^2 \right)^{1/q_k} = \prod_{\ell \in \mathcal{S}_1(\xi_1)} \left(\xi_1 g_{1,\ell}^2 \right)^{1/q_1}, \ \forall k. \tag{4.26}$$

Considering these equations and Eq. 4.25, ξ_k, $k = 1, \ldots, K$, can be viewed as monotonically increasing functions of ξ_1. In the light of Eq. 4.25, this necessarily implies that powers $P_{k,\ell}$ are monotonically increasing functions of ξ_1. As a consequence, for a fixed maximum transmit power P the function

$$f(\xi_1) = P - \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} P_{k,\ell} \tag{4.27}$$

is monotonically decreasing in ξ_1. Clearly, $f(\xi_1 = 0) = P$ and there exists a number M for which $f(\xi_1 = M) < 0$. As this function is continuous, there must exist $\bar{\xi}_1 \in [0, M]$ such that $f(\xi_1 = M) = 0$. Since the function decreases monotonically, this value must be unique. Once found, this value univocally determines the water level of all other users through Eqs. 4.26. Based on these monotonicity properties, a bisection procedure can be used to determine water levels in the following way. First, the water level of a certain user is arbitrarily chosen, e.g., ξ_1. The water levels of all other users are then determined so that Eqs. 4.26 are satisfied. The total power is subsequently computed using Eqs. 4.25. If the resulting total power is smaller than the available transmit power P the level ξ_1 is increased. Otherwise, the level ξ_1 is decreased. These computations are repeated until the required power is approximately equal to the available power.

4.1.6.3 Numerical results

In this section numerical results are presented that correspond to a multicarrier transmission system with $N = 16$ uncorrelated subcarriers. Fig. 4.9 shows average rate pairs for a spatially uncorrelated Rayleigh distributed broadcast channel with unit-variance channel coefficients. The system parameters are $K = 2$, $t = 4$ and $r_k = 2$. Relative rate requirements have been considered such that $R_1/R_2 = 0.1 \times n$ and $R_2/R_1 = 0.1 \times n$ with $n \in \{1, 3, \ldots, 9\}$. Rate vectors have been plotted for three different SNR values. It can be observed that SESAM almost achieves the performance of the optimum solution in both plots. This is specially true for the range of points achieving the maximum sum rate as well as for points close to the axes. For points in between some rate loss can be noticed. However, in any case this loss is observed to be below 3 % of the optimum rate per user. More noticeable is the gap between the successive encoding approaches and the linear decomposition scheme. Nonetheless, the performance loss due to purely linear interference suppression keeps below 15% of the optimum rates for all simulated points. For the same system parameters but for a spatially correlated channel average rate pairs are shown in Fig. 4.10. The correlation is modeled by a covariance matrix with an eigenvalue profile as given by Eq. 4.10. As in previous sections, we observe the general decrease in achievable rate due to correlation. The performance loss of SESAM with respect to the optimum solution is in this case 4.16% for a constraint $R_1/R_2 = 0.1$ at 25 dB. The gap between the linear scheme and the optimum approaches now reaches a maximum of 17.23% at 15 dB for a constraint $R_1/R_2 = 0.9$.

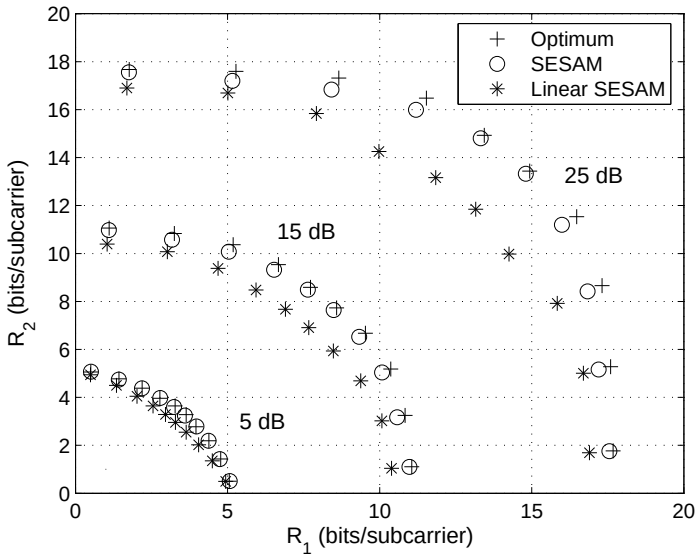

Figure 4.9: Average optimum and suboptimum rate balancing points for an uncorrelated channel with $K = 2$, $t = 4$, $r_k = 2$ and $N = 16$.

Fig. 4.11 shows average rate per user obtained in a broadcast channel with $t = 4$, $r_k = 2$ and a "maximum" fairness constraint, i.e., $q_1 = q_2 = \cdots = q_K$, for $K = 2$, $K = 5$ and $K = 10$ users. On each subcarrier the entries of channel matrices have been indepen-

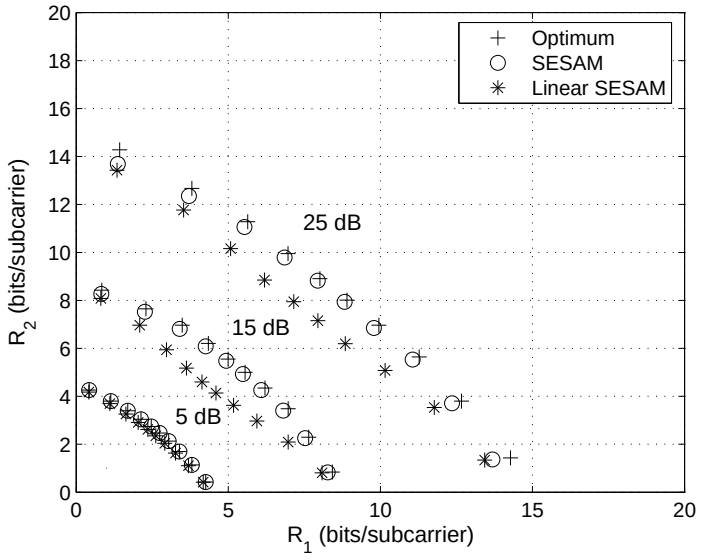

Figure 4.10: Average optimum and suboptimum rate balancing points for a correlated channel with $K = 2$, $t = 4$, $r_k = 2$ and $N = 16$.

dently drawn from a zero-mean complex-valued Gaussian distribution with unit variance. SESAM practically achieves the performance of the optimum solution for 2 and 5 users. By contrast, for the case of 10 users the gap between the optimum solution and SESAM is noticeable. The reason for that might be the high number of users per subcarrier in the system. As the number of users per subcarrier increases, the optimum solution tends to split the users in groups that are served in separate OFDM symbols as part of a time-sharing strategy. By contrast, SESAM tries to comply with the QoS constraint by serving all users simultaneously in each single OFDM symbol. This strategy becomes increasingly inefficient for growing number of users. Although this argument also applies to the linear decomposition scheme, the gap with respect to SESAM narrows as the number of users increases. This is due to the fact that in a system with a large number of users, the probability of allocating nearly orthogonal subchannels on the same subcarrier becomes higher. If subchannels allocated on every subcarrier are mutually orthogonal the gap between the successive-encoding-based decomposition scheme and the linear decomposition scheme disappears.

In Table 4.4 average numbers are given concerning computation and implementation of the optimum solution in Fig. 4.11. In order to compute the optimum solution, the ellipsoid method given in Algorithm 3.11 has been used. In order to solve the weighted sum-rate maximization problem required to update the subgradient of the dual problem at each iteration, Algorithm 3.10 has been applied in conjunction with Algorithm 3.8. As stop condition for the ellipsoid method we require that the maximum radius of the ellipsoid at a certain iteration become smaller than $\epsilon = 0.01$ or, alternatively, that

$$\max_k \left| \frac{R_k^{(\ell)}}{q_k} - \frac{1}{K} \sum_{i=1}^{K} \frac{R_i^{(\ell)}}{q_i} \right| \leq \epsilon,$$

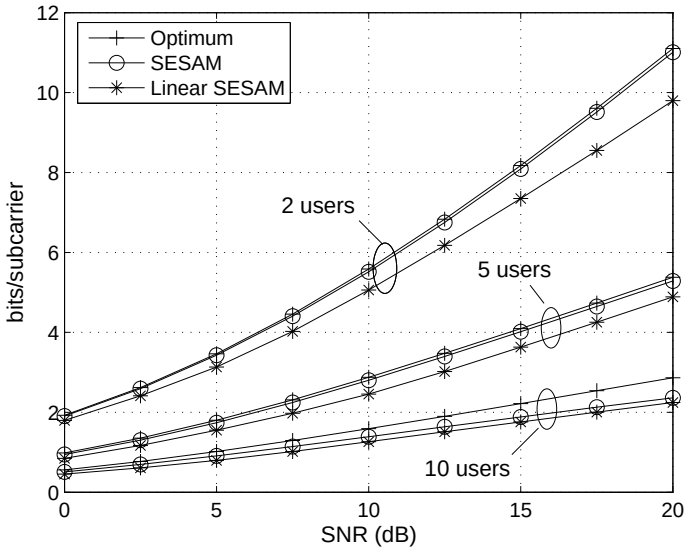

Figure 4.11: Average optimum and suboptimum rates per user with equal rate requirements, i.e., $q_k = 1$, $\forall k$. $N = 16$, $t = 4$ and $r_k = 2$.

i.e., the vector of rates obtained at a certain iteration ℓ is almost parallel to the constraint vector $\boldsymbol{q}$. As stop conditions for Algorithms 3.10 and 3.8, we require that the increment in the value of the respective objective function at a certain iteration to be smaller than 1% and 0.1% of the value achieved in the previous iteration, respectively. Outer iterations refers now to the number of iterations performed by Algorithm 3.11. Inner iterations refers to the average number of iterations performed by Algorithm 3.8 per subcarrier and iteration of Algorithm 3.10. Specially significant is the degradation in convergence speed of the ellipsoid method as the number of users increases (cf. Fig. 3.3). Beside this computational complexity time-sharing poses an additional difficulty to practical implementation. Indeed, having to switch between different transmission strategies increases signaling overhead and the time needed to effectively realize nearly error-free transmission at the desired rates. In Table 4.4 we observe that, for a "maximum" fairness constraint, the average number of necessary time-sharing points approaches the actual number of users.

SNR (dB)	0	5	10	15	20
inner iterations	11.3/17.8/19.9	9.9/19.2/21.7	7.0/17.8/20.3	4.0/15.3/18.0	2.5/13.2/16.0
outer iterations	6.4/85.7/440.6	6.8/85.1/437.4	7.0/84.3/422.2	7.0/83.8/418.4	7.0/83.3/411.6
time-sharing points	1.4/3.1/7.9	1.6/4.4/7.9	1.9/4.5/8.1	2.0/4.7/8.4	2.0/4.7/8.8

Table 4.4: Average numbers involved in the computation and implementation of the optimum rate balancing with fairness QoS constraint for $N = 16$, $t = 4$, $r_k = 2$ and $K = 2/5/10$.

4.2 SINR-based successive subchannel allocation method

In this section an algorithm is presented that, as SESAM, performs a successive allocation of subchannels, but, different from SESAM and all other decomposition methods for that matter, does not impose any zero-forcing constraints on the selection of subchannels. Rather than using channel gains as criterion for the assignment of new subchannels, the algorithm establishes at each step a new subchannel based on an SINR criterion. Note that this is only possible if the allocation of power and dimensions is done in parallel. This is in contrast to the decomposition approaches described above, where dimensions are first allocated and power allocation is carried out in a second stage. A further difference of this algorithm with respect to decomposition approaches consists in the fact that it is applied to the dual MAC of a given broadcast channel rather than to the broadcast channel itself. Subsequently, streamwise duality (cf. Section A.4) can be used in order to find the streamwise strategy that achieves the same stream rates in the broadcast channel. If applied to MIMO OFDM channels, the algorithm can be run in parallel on each of the subcarriers by assuming a uniform power allocation across subcarriers. For this reason and in order to simplify notation, in the sequel, the general MIMO broadcast channel model given in Eq. 2.6 and, in particular, its dual MAC given in Eq. 2.25 are considered. In the next sections we discuss the application of this approach to the sum-rate and the weighted sum-rate maximization problems. Due to the fact that the resulting subchannels are coupled, subsequent optimization of the power allocation in order to enforce a rate balancing constraint (cf. Section 4.1.6.2) is a hardly tractable task. This somehow compromises applicability of this approach to the rate balancing problem.

4.2.1 Sum-rate maximization

Consider allocation of the first subchannel in the MAC and assume that information sent over this subchannel is to be decoded last.[9] Let the first subchannel be allocated to user k. The maximum achievable SINR that can be achieved can be computed as

$$\mathrm{SINR}_k = \max_{\boldsymbol{b}} P\boldsymbol{b}^H \boldsymbol{H}_k \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{b}, \tag{4.28}$$

subject to $\|\boldsymbol{b}\|_1 = 1$, where P is the total transmit power available. SINR_k is achieved by transmitting with all the available power over the stream defined by the beamforming vector $\boldsymbol{b}_k$ that maximizes Problem 4.28. At the receiver, the matched filter or scaled MMSE filter vector $\boldsymbol{a}_k = \alpha_k \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{b}_k$ is applied, where α is an arbitrary scaling factor. This receive filter is an SINR maximizer. Since the maximum SINR achievable on a first allocated subchannel is determined by the beamforming vector $\boldsymbol{b}_k$ and the power P, we can write $\mathrm{SINR}_k(\boldsymbol{b}_k, P)$. The transmit beamforming vector characterizing the stream with highest SINR_k among all users is chosen to represent the first allocated subchannel, i.e.,

$$\boldsymbol{u}_1 = \arg \max_{\boldsymbol{b}_k} \max_k \mathrm{SINR}_k(\boldsymbol{b}_k, P).$$

[9] Due to duality, the corresponding stream in the BC will be encoded in the first place (cf. Section A.4.2).

The maximum achievable rate after allocation of the first subchannel is given by

$$R^{(1)} = \log_2\left(1 + \mathrm{SINR}_{\pi(1)}(\boldsymbol{u}_1, P)\right).$$

Aiming at the allocation of the second subchannel, which will be decoded immediately before the first allocated subchannel, we first divide the power in two equal parts. $P/2$ is assigned to the already established subchannel and $P/2$ is reserved for the subchannel to be assigned in the next step. Now, the maximum SINR achieved by any user k can be written as

$$\mathrm{SINR}_k(\boldsymbol{b}_k, P/2) = \max_{\boldsymbol{b}} \frac{P}{2}\boldsymbol{b}^H \boldsymbol{H}_k \left(\boldsymbol{I}_t + \frac{P}{2}\boldsymbol{H}_{\pi(1)}^{\mathrm{H}}\boldsymbol{u}_1\boldsymbol{u}_1^{\mathrm{H}}\boldsymbol{H}_{\pi(1)}\right)^{-1} \boldsymbol{H}_k^{\mathrm{H}}\boldsymbol{b},$$

subject to $\|\boldsymbol{b}\|_1 = 1$. In order to achieve $\mathrm{SINR}_k(\boldsymbol{b}_k, P/2)$, $\boldsymbol{b}_k$ must be used as beamforming vector at transmitter k. At the receiver, the scaled MMSE filter

$$\boldsymbol{a}_k = \alpha_k \left(\boldsymbol{I}_t + \frac{P}{2}\boldsymbol{H}_{\pi(1)}^{\mathrm{H}}\boldsymbol{u}_1\boldsymbol{u}_1^{\mathrm{H}}\boldsymbol{H}_{\pi(1)}\right)^{-1} \boldsymbol{H}_k^{\mathrm{H}}\boldsymbol{b}_k,$$

must be applied. The transmit beamforming vector characterizing the stream with highest SINR_k is chosen to represent the second allocated subchannel, i.e.,

$$\boldsymbol{u}_2 = \arg\max_{\boldsymbol{b}_k} \max_k \mathrm{SINR}_k(\boldsymbol{b}_k, P/2).$$

The rate obtained after allocation of the two first subchannels can be computed as

$$R^{(2)} = \log_2\left(1 + \mathrm{SINR}_{\pi(1)}(\boldsymbol{u}_1, P/2)\right) + \log_2\left(1 + \mathrm{SINR}_{\pi(2)}(\boldsymbol{u}_2, P/2)\right),$$

where $\mathrm{SINR}_{\pi(1)}(\boldsymbol{u}_1, P/2)$ is the SINR value obtained over the first allocated subchannel after allocation of the first two subchannels. At this point, we compare $R^{(2)}$ and $R^{(1)}$. If $R^{(1)}$ is larger than $R^{(2)}$ the last allocated subchannel is dismissed and allocation is declared completed. If $R^{(2)}$ is larger than $R^{(1)}$ allocation of a third subchannel is pursued. In such case, the total power is divided into three equal parts. Two parts are allocated to the established subchannels and the third part is reserved for the new subchannel. The allocation process proceeds along the lines of the procedure followed for the allocation of the first two subchannels. The pseudocode of this allocation method is given in Algorithm 4.4. Assume that the algorithm terminates as allocation of the $(L+1)$th subchannel results in a decrease of sum rate. In such case, the output of the algorithm consists of the L first allocated subchannels over which the power is uniformly distributed. A streamwise approach achieving the same sum-rate in the dual BC can easily be found by considering DPC-based successive encoding of the streams in the order in which the respective subchannels were allocated. As receive beamforming vector for the ℓth stream, $\boldsymbol{u}_\ell$ should be used. As transmit beamforming vector, the unit-norm scaled MMSE filter achieving $\mathrm{SINR}_{\pi(\ell)}(\boldsymbol{u}_\ell, P/L)$ should be applied. The power allocation can be obtained by solving the linear system of equations given in Eq. A.12.

It is interesting to observe that, if applied to a single-user setting, this algorithm also chooses the singular vectors of the channel as beamforming vectors and performs a kind

Algorithm 4.4 Successive maximum SINR subchannel allocation

1: $\mathrm{SINR}_k(\boldsymbol{b}_k, P) \leftarrow \max_{\boldsymbol{b}} P\boldsymbol{b}^H \boldsymbol{H}_k \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{b}, \quad k = 1, \ldots, K$
subject to $\|\boldsymbol{b}\|_1 = 1$
2: $\boldsymbol{u}_1 \leftarrow \arg\max_{\boldsymbol{b}_k} \max_k \mathrm{SINR}_k(\boldsymbol{b}_k, P)$
3: $R^{(1)} \leftarrow \log_2\left(1 + \mathrm{SINR}_{\pi(1)}(\boldsymbol{u}_1, P)\right)$
4: $\ell \leftarrow 1$
5: **repeat**
6: $\quad \ell \leftarrow \ell + 1$
7: $\quad \mathrm{SINR}_k(\boldsymbol{b}_k, P/\ell) \leftarrow \max_{\boldsymbol{b}} \frac{P}{\ell} \boldsymbol{u}^H \boldsymbol{H}_k \left(\boldsymbol{I}_t + \sum_{i=1}^{\ell-1} \frac{P}{\ell} \boldsymbol{H}_{\pi(i)}^{\mathrm{H}} \boldsymbol{u}_i \boldsymbol{u}_i^{\mathrm{H}} \boldsymbol{H}_{\pi(i)}\right)^{-1} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{b}$,
$\quad k = 1, \ldots, K$, subject to $\|\boldsymbol{b}\|_1 = 1$
8: $\quad \boldsymbol{u}_\ell \leftarrow \arg\max_{\boldsymbol{b}_k} \max_k \mathrm{SINR}_k(\boldsymbol{b}_k, P/\ell)$
9: $\quad R^{(\ell)} \leftarrow \sum_{i=1}^{\ell} \log_2\left(1 + \mathrm{SINR}_{\pi(i)}(\boldsymbol{u}_i, P/\ell)\right)$
10: **until** $R^{(\ell)} < R^{(\ell-1)}$

of quantized waterfilling power loading. This can be seen as follows. Letting $\boldsymbol{H}$ be the channel matrix of the single-user channel and λ_j, $j = 1, \ldots, \mathrm{Rank}\{\boldsymbol{H}\}$, the singular values of this matrix ordered such that $\lambda_j \geq \lambda_{j+1}$, the first allocation step delivers $\mathrm{SINR}_1 = P\lambda_1^2$ and the left singular vector associated with λ_1^2 as vector $\boldsymbol{u}_1$. The matrix

$$\boldsymbol{H}\left(\boldsymbol{I}_t + \frac{P}{2}\boldsymbol{H}^{\mathrm{H}}\boldsymbol{u}_1\boldsymbol{u}_1^{\mathrm{H}}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^{\mathrm{H}}$$

used in order to determine allocation of the second subchannel is easily shown to have the same eigenvectors as $\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}}$ and eigenvalues $\mu_1 = \lambda_1^2/(1 + P\lambda_1^2/2)$, $\mu_j = \lambda_j^2$, $j = 2, \ldots, \mathrm{Rank}\{\boldsymbol{H}\}$. Correspondingly, the beamforming vector selected in this second step is again a left singular vector of $\boldsymbol{H}$. This, in turn, implies that the vectors eligible as beamforming vectors in the next round are again the left singular vectors of $\boldsymbol{H}$ and so on. If $\mu_1 > \mu_2$, $\boldsymbol{u}_2 = \boldsymbol{u}_1$, i.e., the algorithm chooses to transmit two information streams over the same physical spatial dimension. Otherwise, $\boldsymbol{u}_2$ is chosen to be the eigenvector associated with λ_2^2. In the first case, the eigenvalues of the matrix used in order to determine the allocation of the third subchannel are $\mu_1 = \lambda_1^2/(1+2P\lambda_1^2/3)$, $\mu_j = \lambda_j^2$, $j = 2, \ldots, \mathrm{Rank}\{\boldsymbol{H}\}$. In the second case, $\mu_1 = \lambda_1^2/(1+P\lambda_1^2/3)$, $\mu_2 = \lambda_1^2/(1+P\lambda_1^2/3)$, $\mu_j = \lambda_j^2$, $j = 3, \ldots, \mathrm{Rank}\{\boldsymbol{H}\}$. Following this process, it can be easily verified that after allocation of the ℓth subchannel, the total power allocated for transmission over the spatial dimension corresponding to the singular value λ_j is given by $P_j = n_j P/\ell$ where $n_j \in \{0, \ldots, \ell\}$ corresponds to the number of streams assigned for transmission over that dimension. It can also be shown that $n_1 \geq n_2 \geq \ldots \geq n_J$, where $J = \mathrm{Rank}\{\boldsymbol{H}\}$. This example also makes clear an essential difference between subchannels allocated by decomposition approaches and the subchannels allocated by this algorithm. In the first case, subchannels can be identified with spatial dimensions. In fact, the number of subchannels that can be allocated by decomposition approaches is limited to the spatial rank of the system as a whole. By contrast, in the

context of this algorithm, subchannel is synonym of the term information stream. In fact, as we have seen several subchannels or streams can be transmitted over the same spatial dimension. That is, the number of streams or subchannels allocated is not constrained by the dimensionality of the system. Of course, for the single-user setting, in a practical implementation, all streams assigned to one particular dimension can be merged into a single stream, thus simplifying detection.

Fig. 4.12 shows average sum-rate curves for the settings used in Figs. 4.1 and 4.2. Though numerically, it is observed that the SINR-based successive subchannel allocation outperforms SESAM, this is certainly not visible in this plot. Fig. 4.13 compares the performance of SESAM and the SINR-based successive subchannel allocation scheme for the settings used in Figs. 4.3 and 4.4. Only a negligible gap can be appreciated between the curves corresponding to both approaches. This gap seems to be actually larger than that between the optimum approach and SESAM in Figs. 4.3 and 4.4. This is surely due to the numerical error incurred by stopping the search in the iterative optimum algorithm after a finite number of iterations.

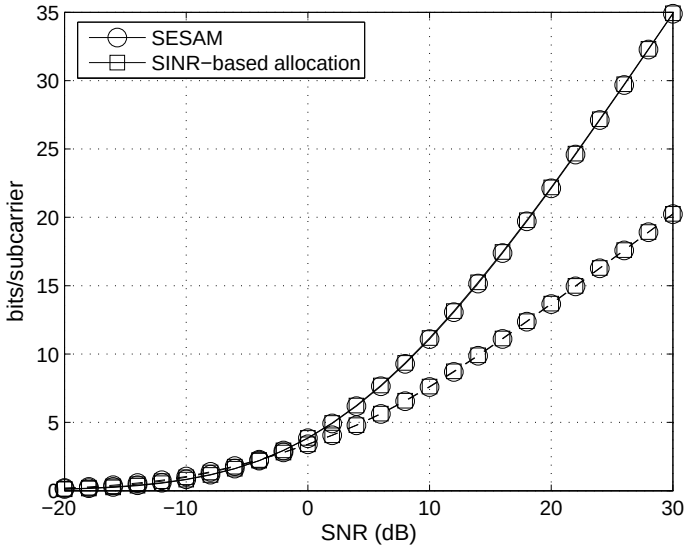

Figure 4.12: Comparison of SINR-based and SESAM sum-rate maximizing allocation for spatially uncorrelated (solid lines) and spatially correlated (dashed lines) channels. $K = 2$, $t = 4$, $r_k = 2$ and $N = 16$.

4.2.2 Weighted sum-rate maximization

Algorithm 4.4 can be readily endowed with a mechanism that considers priorities assigned to users in the allocation process. Surely the easiest way of doing this consists in weighting the SINR values in step 8 with the priorities of the respective users. The pseudocode of the resulting allocation method is given in Algorithm 4.5. If all weights are equal, i.e., $\mu_1 = \cdots = \mu_K$, both Algorithm 4.4 and Algorithm 4.5 yield the same result. If there is only one user with a weight different from zero, the algorithm simply performs an SVD

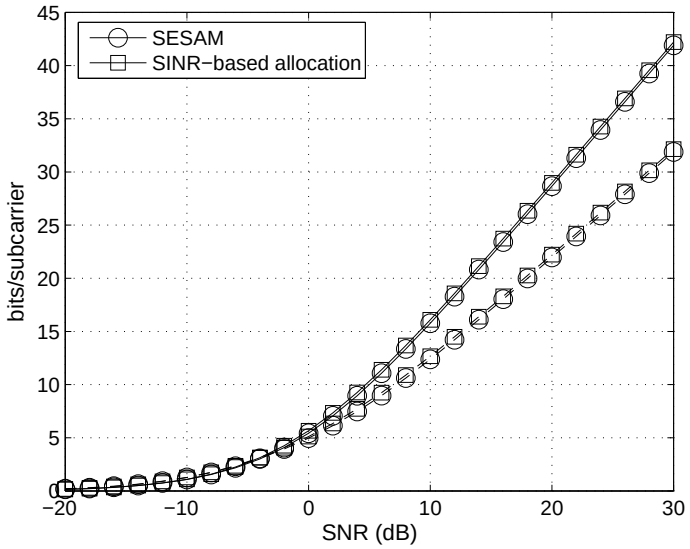

Figure 4.13: Comparison of SINR-based and SESAM sum-rate maximizing allocation for spatially uncorrelated (solid lines) and spatially correlated (dashed lines) channels. $K = 10$, $t = 4$, $r_k = 2$ and $N = 16$.

of the corresponding single-user channel and a quantized waterfilling power allocation as discussed above.

While the weighted sum-rate maximizing SESAM algorithm delivers a set of subchannels over which a priority-sensitive power allocation can be subsequently performed, Algorithm 4.5 delivers a set of coupled subchannels with a power allocation that does not consider priorities in an explicit way. This is a qualitative difference between these two approaches. Figs. 4.14 and 4.15 show a quantitative comparison between both schemes for the same settings used in Figs. 4.5 and 4.6, respectively. In general, the points obtained from the SINR-based approach lie on the same curved line as those of SESAM, however the distribution is different. Points corresponding to the SINR-based scheme are, at least for the uncorrelated channel, not as evenly distributed as those of the SESAM scheme. For the correlated channel, performance loss of the SINR-based scheme with respect to SESAM is especially visible for the points close to the axes at SNR = 25 dB. At these points, for most of the channel realizations only the user with priority 0.9 is served. However, due to the strong correlation this user has two uneven dimensions. While the waterfilling power allocation performed on the top of SESAM adapts to this channel structure optimally, the somehow more rigid quantized power allocation of the SINR-based algorithm turns out to be quite inefficient for some channel realizations. Figs. 4.16 and 4.17 compare both approaches for the settings of Figs. 4.7 and 4.8. While the different in distribution is visible, the SINR-based allocation scheme reaches a weighted sum rate that is 2.70 % below the value achieved by SESAM in Fig. 4.16 and 2.63 % below the value achieved by this approach in Fig. 4.17.

Algorithm 4.5 Successive maximum weighted SINR subchannel allocation

1: $\mathrm{SINR}_k(\boldsymbol{b}_k, P) \leftarrow \max_{\boldsymbol{b}} P\boldsymbol{b}^H \boldsymbol{H}_k \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{b}, \quad k = 1, \ldots, K$
subject to $\|\boldsymbol{b}\|_1 = 1$

2: $\boldsymbol{u}_1 \leftarrow \arg\max_{\boldsymbol{b}_k} \max_k \mu_k \mathrm{SINR}_k(\boldsymbol{b}_k, P)$

3: $R^{(1)} \leftarrow \mu_{\pi(1)} \log_2 \left(1 + \mathrm{SINR}_{\pi(1)}(\boldsymbol{u}_1, P)\right)$

4: $\ell \leftarrow 1$

5: **repeat**

6: $\quad \ell \leftarrow \ell + 1$

7: $\quad \mathrm{SINR}_k(\boldsymbol{b}_k, P/\ell) \leftarrow \max_{\boldsymbol{b}} \frac{P}{\ell} \boldsymbol{u}^H \boldsymbol{H}_k \left(\boldsymbol{I}_t + \sum_{i=1}^{\ell-1} \frac{P}{\ell} \boldsymbol{H}_{\pi(i)}^{\mathrm{H}} \boldsymbol{u}_i \boldsymbol{u}_i^{\mathrm{H}} \boldsymbol{H}_{\pi(i)}\right)^{-1} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{b}$,
$\quad k = 1, \ldots, K$, subject to $\|\boldsymbol{b}\|_1 = 1$

8: $\quad \boldsymbol{u}_\ell \leftarrow \arg\max_{\boldsymbol{b}_k} \max_k \mu_k \mathrm{SINR}_k(\boldsymbol{b}_k, P/\ell)$

9: $\quad R^{(\ell)} \leftarrow \sum_{i=1}^{\ell} \mu_{\pi(\ell)} \log_2 \left(1 + \mathrm{SINR}_{\pi(i)}(\boldsymbol{u}_i, P/\ell)\right)$

10: **until** $R^{(\ell)} < R^{(\ell-1)}$

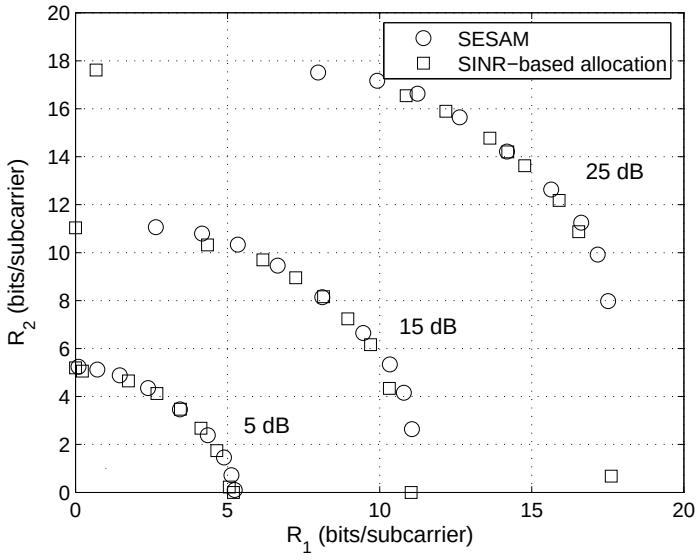

Figure 4.14: Comparison of SINR-based and SESAM weighted sum-rate maximizing allocation for a spatially uncorrelated broadcast channel with $K = 2$, $t = 4$, $r_k = 2$ and $N = 16$.

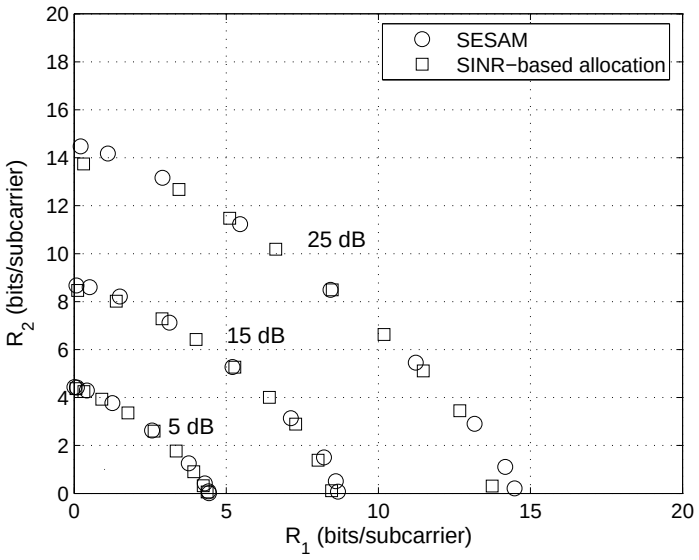

Figure 4.15: Comparison of SINR-based and SESAM weighted sum-rate maximizing allocation for a spatially correlated broadcast channel with $K = 2$, $t = 4$, $r_k = 2$ and $N = 16$.

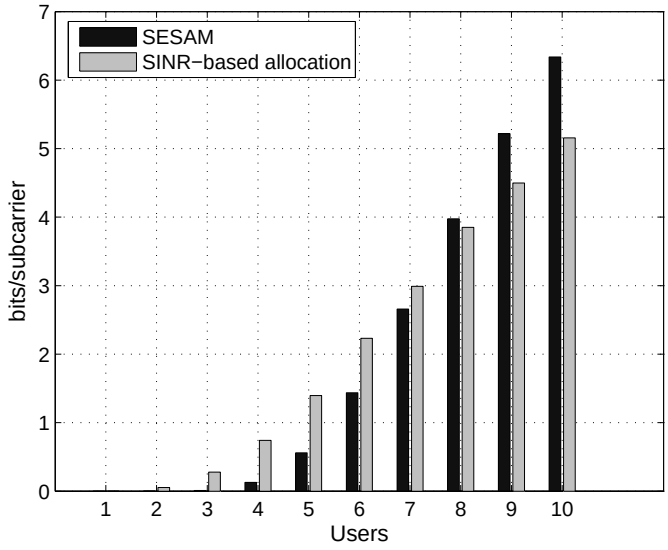

Figure 4.16: Comparison of SINR-based and SESAM weighted sum-rate maximizing allocation for a spatially uncorrelated broadcast channel with $K = 10$, $t = 4$, $r_k = 2$ and $N = 16$. SNR = 15 dB.

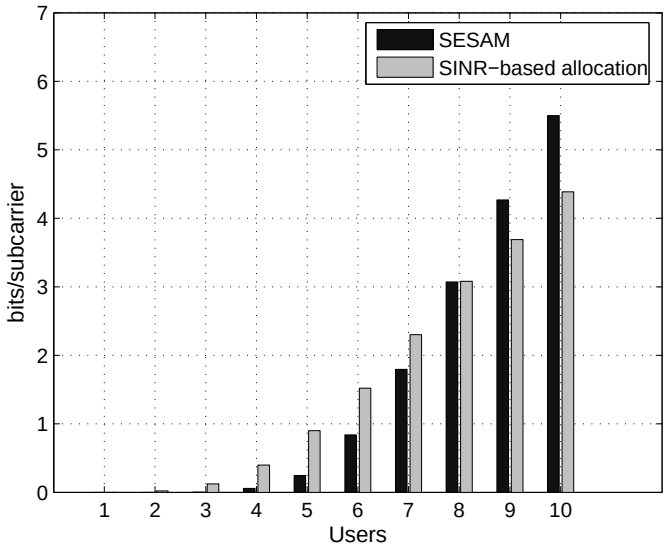

Figure 4.17: Comparison of SINR-based and SESAM weighted sum-rate maximizing allocation for a spatially correlated broadcast channel with $K = 10$, $t = 4$, $r_k = 2$ and $N = 16$. SNR = 15 dB.

5 Feedback of channel state information

5.1 Delay-limited and rate-limited feedback paradigms

Recently, feedback of channel state information has attracted the interest of many researchers in the communications area. Most of the work done so far follows an approach that can be called the rate-limited feedback paradigm. This approach is based on the assumption of an error-free feedback channel over which a limited number of bits is transmitted that convey channel state information from a receiver to the transmitter (e.g., [84, 78, 154, 64]). A comprehensive overview on this body of research can be found in [77]. Typically, a major goal of these publications consists of analyzing the impact of the quantization error on a performance measure in the forward link, such as capacity or outage probability. Due to the error-free assumption, transmission of the quantization bits in the feedback link is not an issue. From a theoretical point of view, zero error probability is strictly impossible if transmission is carried out over a finite number of channel uses in the feedback link, i.e., zero error probability requires an unbounded transmission delay. From a practical point of view, however, the probability of error can realistically be neglected if the the number of admissible channel uses is large enough relative to the number of transmitted feedback bits, and enough diversity is available in the feedback link. In simple communication systems, such as single-user, single-carrier MISO or MIMO links, a few feedback bits have been shown to be enough to closely approximate perfect CSIT performance [78, 95]. By contrast, the number of feedback bits required to approximate full channel knowledge performance appears to be considerable larger in single-user MIMO-OFDM systems (cf. [29, 88]) and can become even larger in multiuser systems, where it must be scaled by the number of transmit antennas and the transmit power in order to maintain an SNR-independent performance loss [64, 91]. Obviously, almost error-free transmission of a large number of feedback bits can only be assumed at the cost of considerable delays. Depending on the rate of change of the forward channel, these delays might render the CSI obsolete or become an unaffordable overhead for the communication system. Besides, fading feedback channels with a low degree of spatial and frequency diversity make the occurrence of transmission errors unavoidable as, due to delay limitations, time diversity will generally be impossible to leverage. As a consequence, at least in multiuser systems with time dispersive forward channels and fading feedback channels, it seems appropriate to shift in the way of addressing the CSI feedback problem from the rate-limited paradigm followed so far to a more realistic delay-limited paradigm, where parameters such as the number of feedback bits are left open and, instead, a strict delay limitation is imposed [111, 112].

While feedback of channel state information under the rate-limited paradigm can be viewed as a source coding problem [154], under the delay-limited paradigm it can be viewed as a joint source and channel coding problem, which is notably more complex. The source is represented by the channel coefficients of the forward link that must be adequately

encoded in order to be transmitted over a given number of channel uses in the feedback link. The goal is to minimize a distortion measure related to performance in the forward link. For this problem, optimality of digital transmission is stated by the joint source and channel coding theorem [101, Theorem 21], also known as separation theorem, provided that the lengths of the codewords tend to infinity. However, under a strict delay limitation there is not theoretical evidence of the optimality of digital transmission. Lately, linear analog approaches have been proposed by several authors [81, 96, 120, 131] for feedback purposes. These schemes are appealing due to their simplicity, the transmitter and receiver consisting of a linear precoder and a linear filter, respectively. This is a significant advantage with respect to high-performance digital approaches including vector quantizers, channel coding and detection algorithms. Due to the complexity of the issue, this chapter pursues the humble but non-trivial goal of providing a basic understanding of the fundamental difference between linear analog approaches and delay-limited digital approaches when employed for feedback of channel state information in wireless communication systems. To this end, first, a simple SISO OFDM fading feedback link model is assumed in the next section. Based on this model and considering a mean squared error (MSE) distortion measure, some analysis will be carried out concerning the performance of analog and digital approaches and theoretical bounds. In Section 5.3, part of this analysis will be extended to a feedback link with multiple antennas. Finally, in Section 5.4, performance in the forward link is considered and numerical results are shown that illustrate the impact of both analog and digital feedback transmission schemes. For a specific single-carrier multiuser setting, there has been some recent work on the comparison of digital and linear analog transmission approaches reported in [19, 20].

5.2 Single-input single-output time-dispersive fading feedback channel

5.2.1 Feedback link model

The block diagram of the feedback link that will be considered in this section is given in Fig. 5.1. At regular time intervals the source delivers a vector $\boldsymbol{h} \in \mathbb{C}^L$ as an output, whose entries are assumed to be statistically independent and distributed according to a zero-mean circularly symmetric Gaussian distribution with unit variance. The source is assumed to be memoryless, meaning that outputs at different time instants are statistically independent. The encoder is a function that maps a source output into a multivariate signal $\boldsymbol{w} \in \mathbb{C}^N$. Here, we assume that the dimensionality of the source output is equal to or smaller than the dimensionality of the signal space, i.e., $L \leq N$. To the vector of transmit signals an average power constraint applies, i.e.,

$$\mathrm{E}\left\{\|\boldsymbol{w}\|_2^2\right\} \leq P. \tag{5.1}$$

The transmit signal is transformed by a diagonal channel matrix $\boldsymbol{G} \in \mathbb{C}^{N \times N}$ and to the resulting signal the noise vector $\boldsymbol{n} \in \mathbb{C}^N$ is added, i.e.,

$$\boldsymbol{r} = \boldsymbol{G}\boldsymbol{w} + \boldsymbol{n}.$$

Entries g_n, $n = 1, \ldots, N$, on the main diagonal of $\boldsymbol{G}$ are assumed to be realizations of a zero-mean circularly symmetric Gaussian distribution with unit variance. For purposes of analysis, a block-fading model is considered. The channel coefficients $g_{1,\ldots,N}$ are divided in M blocks comprising N/M consecutive coefficients. Coefficients of different blocks are statistically independent. Coefficients in the same block are fully correlated, i.e., their realizations are identical. The additive noise vector is assumed to be zero-mean, white, circularly symmetric Gaussian with unit variance per complex dimension. It is assumed that the encoder does not know matrix $\boldsymbol{G}$ but only its statistics. The decoder is a function that maps the received signal $\boldsymbol{r} \in \mathbb{C}^N$ into an estimate of the source output $\hat{\boldsymbol{h}} \in \mathbb{C}^L$. The decoder is assumed to know matrix $\boldsymbol{G}$.

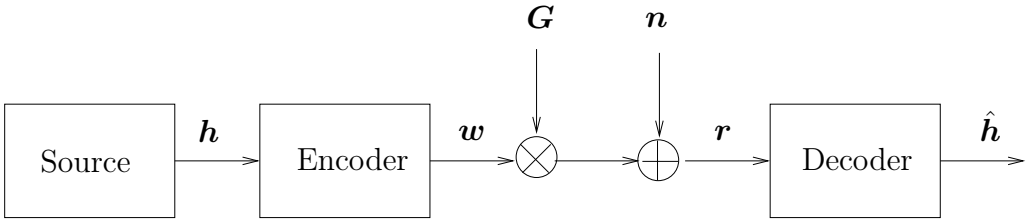

Figure 5.1: Feedback link model.

At a given time, the output of the source represents information about the instantaneous state of the channel in the forward link. Several kinds of channels can be thought of that suit the assumptions made for the source. If we assume that the channel state is fed back once per coherence time and the forward link is a Rayleigh-fading time-dispersive channel with L uncorrelated taps and a flat power delay profile, the entries h_ℓ, $\ell = 1, \ldots, L$, of the source output $\boldsymbol{h}$ might represent each of the forward channel delay taps. Alternatively, these coefficients might represent the coefficients of a MIMO matrix with L entries corresponding to a non-dispersive Rayleigh-fading uncorrelated MIMO forward channel. Any combination of these two models with uncorrelated antennas and taps, and flat power delay profiles can be accommodated into this source model. If feedback is performed at intervals shorter than the coherence time of the forward channel, correlations would exist between the channel states observed in two consecutive channel measurements. In this case, the outputs of the source could be interpreted as the innovations of the new estimate with respect to previous estimates, which do not exhibit any temporal correlation (cf. [131]). The feedback channel model also accepts several interpretations. For instance, it can be viewed to represent a SISO OFDM channel, being g_n the complex-valued channel gain of the nth subcarrier. Alternatively, it can be viewed as a sequence of N uses of a non-dispersive SISO channel. A more general way to look at this model is as a sequence of uses of a SISO OFDM channel with the product of channel uses and the number of subcarriers being N.

While the source and the channel in the above model are assumed to be given and fixed, the encoder and decoder are the blocks that can be designed in order to optimize a performance measure of interest. Here, as already suggested by Fig. 5.1 and in the above description, the goal is to reproduce the source output as faithfully as possible at the output of the decoder. To this end, we choose the mathematically convenient mean squared error

(MSE) as a figure of merit,

$$\epsilon = \frac{1}{L}\mathrm{E}\{\|\boldsymbol{h} - \hat{\boldsymbol{h}}\|_2^2\}. \tag{5.2}$$

Setting this as a goal, fully decouples the design of the feedback link from the design of the transmission strategy in the forward link. On the one hand, the results obtained using this figure of merit have a general character since they are independent of the particularities of the forward link. On the other hand, for a particular forward link, a feedback link designed according to the MSE figure of merit will, in general, provide suboptimum results in terms of performance achievable in the forward link. Indeed, the feedback link (encoder and decoder) should be ideally designed so that at the output of the decoder some parameters are delivered that define the best transmission strategy for the forward link according to a performance measure of interest. However, even for simple forward channels, the resulting design problem turns out extraordinarily difficult. By contrast, an MSE figure of merit, though suboptimum in terms of forward link performance, seems to be correlated with any of the usual performance measures in the forward link and, at the same time, allows the analysis of some fundamental questions concerning transmission of channel state information over the feedback link.

5.2.2 Theoretical upper bounds

5.2.2.1 Optimum performance theoretically achievable

There are two kinds of temporal constraints that apply to the feedback link model described in the previous section. The first is due to the finite dimensionality of matrix $\boldsymbol{G}$, whose entries represent the channel gains of a finite number of channel uses. The second is that the estimate at the output of the decoder at a given time instant depends causally on the sequence of source outputs. That is, if $\hat{\boldsymbol{h}}[1], \hat{\boldsymbol{h}}[2], \ldots$ is the sequence of estimates at the output of the decoder and $\boldsymbol{h}[1], \boldsymbol{h}[2], \ldots$ is the sequence of source outputs, $\hat{\boldsymbol{h}}[k]$ depends at most on $\boldsymbol{h}[k]$ and the preceding source outputs but not on subsequent source outputs. In practical terms, this constraint means that, at every time instant, the output of the feedback link contains as much information as possible about the last channel state measured at the receiver of the forward link and not about previously measured channel states. That is, there is no delay caused by buffering of information and the only delay is due to the number of channel uses and duration of the transmit symbols. Ignoring this delay constraint due to causality, in the following, we derive an upper bound on performance for the feedback link that is based on rate-distortion theory. In accordance with existing joint source and channel coding literature, we call this bound optimum performance theoretically achievable (OPTA).

The rate-distortion function of a random variable $\boldsymbol{h}$ with probability density function $p(\boldsymbol{h})$ is defined as

$$R(\epsilon) = \min_{p(\hat{\boldsymbol{h}}|\boldsymbol{h})} \{I(\boldsymbol{h}, \hat{\boldsymbol{h}})\} \quad \text{subject to} \quad \frac{1}{L}\mathrm{E}\{\|\boldsymbol{h} - \hat{\boldsymbol{h}}\|_2^2\} \leq \epsilon, \tag{5.3}$$

where $I(\boldsymbol{h}, \hat{\boldsymbol{h}})$ is the mutual information of variables $\boldsymbol{h}$ and $\hat{\boldsymbol{h}}$, $p(\hat{\boldsymbol{h}}|\boldsymbol{h})$ is the conditional probability density function of $\hat{\boldsymbol{h}}$ given $\boldsymbol{h}$ and an MSE distortion function has been consid-

ered. The rate-distortion function indicates the minimum rate that is necessary in order to encode the source with a distortion no larger than ϵ. For $k \to \infty$, this rate can be approximated as tightly as we want by generating a codebook consisting of 2^{kR} sequences of length k drawn at random according to $p(\hat{\boldsymbol{h}})$ and selecting for each source sequence $\boldsymbol{h}[1], \boldsymbol{h}[2], ..., \boldsymbol{h}[k]$ a jointly typical code vector $\hat{\boldsymbol{h}}[1], \hat{\boldsymbol{h}}[2], ..., \hat{\boldsymbol{h}}[k]$. Roughly speaking, the existence of such a typical code vector is guaranteed by the fact that $R > I(\boldsymbol{h}, \hat{\boldsymbol{h}})$ (see Appendix A.2). As the source sequence and the corresponding code vector are jointly typical, by the law of large numbers, the distortion constraint is also satisfied [40].

According to the channel coding theorem, assigning an index $i \in \{1, 2, \ldots, 2^{kR}\}$ to every code vector, these can be transmitted without error over a channel with capacity $C > R$ provided that $k \to \infty$. At the receiver, an index can be mapped back to the corresponding code vector that represents the original source sequence with a distortion smaller or equal to ϵ. By contrast, if $C < R$, no error-free transmission is possible. As a consequence, original source sequences and eventually detected code vectors are not any more necessarily jointly typical, thereby causing distortion to increase. Thus, we observe that any distortion ϵ is achievable over a channel with capacity C if $R(\epsilon) < C$ and, conversely, if a distortion ϵ is achievable over a channel with capacity C, then $R(\epsilon) < C$ must hold. This result [101, Theorem 21] is known as joint source-channel coding theorem or separation theorem. The name separation theorem comes from the fact that source coding and channel coding can be performed separately. Indeed, source code vectors can be represented by indexes and these can be arbitrarily mapped to codewords of an optimum channel code book, the design of the source and channel codebooks being done independently of each other. Thus, separation of source and channel coding is a way to achieve minimum distortion. However, as we shall soon see, at least in some cases, this is not the only way. As a consequence of this theorem and the fact that $R(\epsilon)$ is a non-increasing function of ϵ, the minimum distortion achievable over a channel with capacity C can be computed by solving $R(\epsilon) = C$.

For $\boldsymbol{h} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_L)$ the rate-distortion function can be derived by noting the following inequalities:

$$\begin{aligned} I(\boldsymbol{h}, \hat{\boldsymbol{h}}) &= h(\boldsymbol{h}) - h(\boldsymbol{h}|\hat{\boldsymbol{h}}) \\ &= L\log(\pi \mathrm{e}) - h(\boldsymbol{h} - \hat{\boldsymbol{h}}|\hat{\boldsymbol{h}}) && (5.4) \\ &\geq L\log(\pi \mathrm{e}) - h(\boldsymbol{h} - \hat{\boldsymbol{h}}) && (5.5) \\ &\geq L\log(\pi \mathrm{e}) - h(\mathcal{CN}(\boldsymbol{0}, \boldsymbol{R}_\epsilon)) && (5.6) \\ &\geq L\log(\pi \mathrm{e}) - L\log(\pi \mathrm{e}\epsilon) && (5.7) \\ &= L\log\left(\frac{1}{\epsilon}\right). \end{aligned}$$

Eq. 5.4 is a consequence of the fact that translation does not change differential entropy. Eq. 5.5 follows from the fact that conditioning reduces entropy. In Eq. 5.6, $\boldsymbol{R}_\epsilon$ is the error covariance matrix, which must fulfil $\mathrm{Tr}\{\boldsymbol{R}_\epsilon\} \leq L\epsilon$. This inequality follows from the fact that circularly symmetric Gaussian distributed variables are entropy maximizers [119]. Finally, Eq. 5.7 is obtained by noting that $\mathrm{Tr}\{\boldsymbol{R}_\epsilon\}/L$ is the arithmetic mean of the eigenvalues of $\boldsymbol{R}_\epsilon$ and $|\boldsymbol{R}_\epsilon|^{1/L}$ is the geometric mean of these eigenvalues. Using the geometric and

arithmetic mean inequality we can write

$$h(\mathcal{CN}(\mathbf{0}, \boldsymbol{R}_\epsilon)) = L \log \left(\pi \mathrm{e} |\boldsymbol{R}_\epsilon|^{1/L}\right) \leq L \log \left(\pi \mathrm{e} \mathrm{Tr}\{\boldsymbol{R}_\epsilon\}/L\right) \leq L \log(\pi \mathrm{e} \epsilon).$$

If $\epsilon < 1$, choosing $p(\hat{\boldsymbol{h}}|\boldsymbol{h})$ such that $\boldsymbol{h} = \hat{\boldsymbol{h}} + \boldsymbol{n}$ with $\hat{\boldsymbol{h}} \sim \mathcal{CN}\left(0, (1-\epsilon)\boldsymbol{I}_L\right)$ and $\boldsymbol{n} \sim \mathcal{CN}\left(0, \epsilon \boldsymbol{I}_L\right)$, all inequalities become equalities. If $\epsilon \geq 1$, $\hat{\boldsymbol{h}}$ can be chosen to be equal to 0 with probability 1. The distortion achieved in this case equals 1 and the required rate is 0. Summing up, for $\boldsymbol{h} \sim \mathcal{CN}(0, \boldsymbol{I}_L)$ the rate-distortion function is given by

$$R(\epsilon) = \begin{cases} L \log \left(\frac{1}{\epsilon}\right) & \text{if} \quad \epsilon < 1 \\ 0 & \text{if} \quad \epsilon \geq 1 \end{cases}. \tag{5.8}$$

This is a trivial extension to multivariate circularly symmetric Gaussian variables of the computation of the rate distortion function for Gaussian variables described in [40].

The capacity of the feedback channel is given by

$$C = \sum_{n=1}^{N} \mathrm{E}\left\{\log\left(1 + \frac{|\boldsymbol{g}_n|^2 P}{N}\right)\right\} = \frac{N}{\log_\mathrm{e} 2} \mathrm{E}_1\left(\frac{N}{P}\right) \exp\left(\frac{N}{P}\right), \tag{5.9}$$

where $\mathrm{E}_1(\cdot)$ is an exponential integral function.[1] That a uniform power allocation is optimum can be shown by considering the following problem,

$$\max_{P_1,\ldots,N} \sum_{n=1}^{N} \mathrm{E}\left\{\log\left(1 + |\boldsymbol{g}_n|^2 P_n\right)\right\} \quad \text{subject to} \quad \sum_{n=1}^{N} P_n \leq P,$$

and noticing that $P_n = P/N$, $\forall n$, satisfies the KKT conditions, which are sufficient due to convexity in this case.

Equating Eqs. 5.8 and 5.9 and solving for ϵ, we obtain the OPTA as

$$\epsilon_{\mathrm{OPTA}} = \exp\left(-\frac{N}{L} \mathrm{E}_1\left(\frac{N}{P}\right) \exp\left(\frac{N}{P}\right)\right). \tag{5.10}$$

This distortion could be approximated by providing the encoder depicted in Fig. 5.1 with buffering capability in order to store long sequences of source outputs and letting it quantize these sequences and map the reproduction values to codewords of a channel code stretching over many different realizations of the channel matrix $\boldsymbol{G}$ in order to transmit at a rate close to C with negligible error probability.

5.2.2.2 Optimum performance theoretically achievable with limited diversity

An alternative upper bound to the OPTA derived in the previous section is obtained as follows. As a lower bound on the number of bits required to represent ocurrences of

[1] $\mathrm{E}_n(x) = \int_x^\infty \frac{e^{-t}}{t^n} dt$.

the source outputs with an average distortion smaller than ϵ, we still consider the rate-distortion function in Eq. 5.8. Given a fixed feedback channel realization $\boldsymbol{G}$, the maximum number of bits that can be reliably transmitted with uniform power allocation[2] reads

$$C(q_{1,\dots,N}) = \sum_{n=1}^{N} \log_2 \left(1 + q_n P/N\right), \tag{5.11}$$

where $q_n = |g_n|^2$, $n = 1, \dots, N$. Correspondingly, the minimum distortion that could possibly be achieved over a particular channel realization can be computed by considering Eqs. 5.8 and 5.11 and yields

$$\epsilon(q_{1,\dots,N}) = \frac{1}{\prod_{n=1}^{N} \left(1 + q_n P/N\right)^{\frac{1}{L}}}. \tag{5.12}$$

This represents a lower bound on the distortion that can be achieved on the feedback link for the specific feedback channel realization. Thus, averaging this expression over channel realizations gives a lower bound on the average distortion achievable over the fading feedback channel. That the resulting bound is tighter than the OPTA derived in the previous section can be shown by noting

$$\epsilon_{\text{OPTA}-\text{LD}} = \mathrm{E}\{\epsilon(\boldsymbol{q}_{1,\dots,N})\} = \mathrm{E}\{\epsilon(C(\boldsymbol{q}_{1,\dots,N}))\} \geq \epsilon(\mathrm{E}\{C(\boldsymbol{q}_{1,\dots,N})\}) = \epsilon_{\text{OPTA}},$$

where the inequality is due to the fact that the distortion-rate function $\epsilon(R)$ is convex. Note that ϵ_{OPTA} is obtained by transmitting codewords over all possible channel realizations, thereby profiting from an unbounded source of diversity. In the new bound, however, the amount of diversity exploited during transmission is that available in just one channel realization. For this reason, we call this bound OPTA with limited diversity (OPTA-LD). Using the block fading assumption on the statistics of the coefficients of the feedback channel matrix $\boldsymbol{G}$, we can write

$$\epsilon_{\text{OPTA}-\text{LD}} = \left(\mathrm{E}\left\{ \frac{1}{(1 + \boldsymbol{q} P/N)^{\frac{N}{ML}}} \right\} \right)^{M}, \tag{5.13}$$

where $\boldsymbol{q}$ is exponentially distributed with mean value 1. Making use of the identity [87]

$$\int_0^\infty \frac{1}{(a+t)^\nu} e^{-pt} dt = p^{\nu-1} e^{ap} \Gamma(-\nu + 1, ap),$$

the expected value in Eq. 5.13 can be computed as

$$\epsilon_{\text{OPTA}-\text{LD}}^{\frac{1}{M}} = \left(\frac{N}{P}\right)^{\frac{N}{ML}} \exp\left(\frac{N}{P}\right) \Gamma\left(-\frac{N}{ML} + 1, \frac{N}{P}\right), \tag{5.14}$$

where $\Gamma(a, x) = \int_x^\infty t^{a-1} e^{-t} dt$ is the upper incomplete gamma function.

[2]Recall that the transmitter does not have knowledge of $\boldsymbol{G}$

5.2.2.3 Asymptotical analysis

Using the inequalities [1]

$$\frac{1}{2}\log_{\mathrm{e}}(1+2x) \leq E_1\left(\frac{1}{x}\right)\exp\left(\frac{1}{x}\right) \leq \log_{\mathrm{e}}(1+x), \tag{5.15}$$

the following asymptotical behavior can be observed for ϵ_{OPTA},

$$\epsilon_{\mathrm{OPTA}}(\mathrm{dB}) = -\eta\frac{N}{L}P(\mathrm{dB}) + O(1), \quad P \to \infty,$$

with $1/2 \leq \eta \leq 1$. That is, the asymptotic distortion decay rate (DDR), defined as

$$\mathrm{DDR} = \lim_{P\to\infty} -\frac{\log\epsilon}{\log P}, \tag{5.16}$$

is proportional to the bandwidth expansion factor defined as the ratio between the dimension of the channel and the dimension of the source output. In the low SNR regime, substitution of any of the bounds of Eq. 5.15 in Eq. 5.10 and a Taylor expansion around zero yield

$$\epsilon_{\mathrm{OPTA}} = 1 - P/L + o(P). \tag{5.17}$$

That is, the behavior of ϵ_{OPTA} in the low SNR regime does not depend on the dimensionality of the channel but only on the transmit power and the dimensionality of the source outputs.

In order to analyze the asymptotical behavior of Eq. 5.14 at high SNR values, it is convenient to write $\Gamma(a,x)$ as a power series [1],

$$\Gamma(1-\nu,x) = \Gamma(1-\nu) - \frac{x^{1-\nu}}{1-\nu} + x^{1-\nu}\sum_{n=1}^{\infty}\frac{(-x)^n}{(1-\nu+n)n!} \quad \nu \neq 1,2,3,\ldots \tag{5.18}$$

Substituting this expression into Eq. 5.14 with $\nu = N/ML$ and $x = N/P$ we obtain

$$\epsilon_{\mathrm{OPTA-LD}}^{\frac{1}{M}} = \begin{cases} \frac{\Gamma(1-\frac{N}{ML})N^{\frac{N}{ML}}}{P^{\frac{N}{ML}}}\exp\left(\frac{N}{P}\right) + o\left(\left(\frac{1}{P}\right)^{\frac{N}{ML}}\right), & \frac{N}{ML} < 1 \\ \frac{N}{\left(\frac{N}{ML}-1\right)P}\exp\left(\frac{N}{P}\right) + o\left(\frac{1}{P}\right), & \frac{N}{ML} > 1 \end{cases}, \quad P \to \infty.$$

That is, if $L > N/M$, the distortion decay rate is equal to N/L and independent of the amount of statistically independent variables (diversity degree) in the feedback channel. On the other hand, if $L \leq N/M$ the decay rate is equal to M, i.e., it is limited by the amount of diversity available in the channel and independent of the bandwidth expansion factor N/L. In particular, if $M = 1$, i.e., the feedback channel is flat fading, the decay rate of any feedback transmission approach is lower bounded by 1 no matter how large the bandwidth available is or how many uses of the feedback link are made in order to transmit the outputs of the source. If $\frac{N}{ML} = 1,2,3,\ldots$, the series expansion in Eq. 5.18

does not hold. However, the conclusions drawn above are still valid. Indeed, in this case, the expressions

$$\Gamma(1-\nu,x) = x^{1-\nu}E_\nu(x), \tag{5.19}$$

$$\frac{1}{x+\nu} < e^x \mathrm{E}_\nu(x) < \frac{1}{x+\nu-1}, \tag{5.20}$$

$\nu = 1, 2, 3, \ldots$, can be used in order to show

$$\epsilon_{\mathrm{OPTA-LD}}^{\frac{1}{M}} = \begin{cases} \frac{N}{P}\mathrm{E}_1\left(\frac{N}{P}\right)\exp\left(\frac{N}{P}\right), & \nu = 1 \\ \frac{N\xi(P,\nu)}{P}, & \nu = 2, 3, \ldots \end{cases}$$

where $\xi(P,\nu)$ is a function that tends to a constant value greater than zero as $P \to \infty$. Note that the asymptotic distortion decay rate is in any case equal to M.

In order to analyze the behavior of the normalized distortion at low SNR values, we use the continued fraction representation of the incomplete gamma function given by [1]

$$\Gamma(1-\nu,x) = e^{-x}x^{1-\nu}\cfrac{1}{x+\cfrac{\nu}{1+\cfrac{1}{x+\cfrac{1+\nu}{1+\cfrac{2}{x+\cdots}}}}}.$$

Substituting this expression in Eq. 5.14 we obtain

$$\epsilon_{\mathrm{OPTA-LD}} = \left(\frac{1}{1+\xi(P)P}\right)^M,$$

where $\xi(P)$ is a function of P that tends linearly to $1/ML$ as $P \to 0$. Correspondingly, a Taylor expansion around 0 yields

$$\tilde{\epsilon} = 1 - P/L + o(P).$$

Note that this expression coincides with Eq. 5.17. That is, in the low SNR regime, both bounds behave identically.

5.2.3 Analog transmission

We say that the transmission over the feedback link is analog if the encoder is a function that maps the outputs of the source to a non-discrete signal space, which is generally uncountable. Here, we focus on linear analog schemes, which are particularly appealing due to their simplicity.[3] In these kind of schemes, the encoder is a linear mapping represented

[3]For work on analog non-linear mappings see [17, 28, 102].

by a matrix $\boldsymbol{T} \in \mathbb{C}^{N \times L}$. In order to comply with the transmit power constraint given by Eq. 5.1, for this matrix,

$$\mathrm{Tr}\left\{\boldsymbol{T}\boldsymbol{T}^{\mathrm{H}}\right\} \leq P \tag{5.21}$$

must hold. The signal at the input of the decoder can be written as a function of the source output as

$$\boldsymbol{r} = \boldsymbol{G}\boldsymbol{T}\boldsymbol{h} + \boldsymbol{n}.$$

Due to joint Gaussianity of this signal and the source output $\boldsymbol{h}$, the estimator that achieves the minimum distortion is the linear minimum mean squared error (MMSE) filter, which reads

$$\boldsymbol{W} = \left(\boldsymbol{I} + \boldsymbol{T}^{\mathrm{H}}\boldsymbol{G}^{\mathrm{H}}\boldsymbol{G}\boldsymbol{T}\right)^{-1}\boldsymbol{T}^{\mathrm{H}}\boldsymbol{G}^{\mathrm{H}},$$

Accordingly, the estimate is given by

$$\hat{\boldsymbol{h}} = \boldsymbol{W}\boldsymbol{r}.$$

Substituting this estimate in Eq. 5.2 the distortion incurred by this optimum receiver can be written as

$$\epsilon = \mathrm{Tr}\left\{\mathrm{E}\left\{\left(\boldsymbol{I} + \boldsymbol{T}^{\mathrm{H}}\boldsymbol{Q}\boldsymbol{T}\right)^{-1}\right\}\right\}, \tag{5.22}$$

where $\boldsymbol{Q} = \boldsymbol{G}^{\mathrm{H}}\boldsymbol{G}$ and the expected value is taken with respect to this matrix. Minimization of the distortion can be carried out over the choice of $\boldsymbol{T}$ subject to the constraint given in Eq. 5.21, i.e.,

$$\min_{\boldsymbol{T}} \mathrm{Tr}\left\{\mathrm{E}\left\{\left(\boldsymbol{I} + \boldsymbol{T}^{\mathrm{H}}\boldsymbol{Q}\boldsymbol{T}\right)^{-1}\right\}\right\} \quad \text{s. t.} \quad \mathrm{Tr}\left\{\boldsymbol{T}\boldsymbol{T}^{\mathrm{H}}\right\} \leq P. \tag{5.23}$$

Even with a model as simple as the one assumed here, a general solution to this problem is difficult to obtain. We next analyze some particular interesting cases which will provide us with some valuable insights into the problem.

5.2.3.1 Flat Fading Feedback Channel ($M = 1$)

For this particular case, $\boldsymbol{Q} = q\boldsymbol{I}_{N \times N}$ and problem (5.23) simplifies to

$$\min_{\boldsymbol{T}} \mathrm{E}\left\{\mathrm{Tr}\left\{\left(\boldsymbol{I} + q\boldsymbol{T}^{\mathrm{H}}\boldsymbol{T}\right)^{-1}\right\}\right\} \quad \text{s. t.} \quad \mathrm{Tr}\left\{\boldsymbol{T}\boldsymbol{T}^{\mathrm{H}}\right\} \leq P.$$

Let $\tau_{1,\ldots,L}$ be the eigenvalues of the product $\boldsymbol{T}^{\mathrm{H}}\boldsymbol{T}$. In terms of these eigenvalues, the optimization problem can be now rewritten as

$$\min_{\tau_{1,\ldots,L}} \mathrm{E}\left\{\sum_{\ell=1}^{L}\left(1 + q\tau_\ell\right)^{-1}\right\} \quad \text{s. t.} \quad \sum_{\ell=1}^{L}\tau_\ell \leq P. \tag{5.24}$$

That is, the solution does exclusively depend on the singular values of the precoder $\boldsymbol{T}$ and not on its singular vectors. In other words, the singular vectors can be arbitrarily chosen. For a particular realization of q the solution of

$$\min_{\tau_{1,\ldots,L}} \sum_{\ell=1}^{L}\left(1 + q\tau_\ell\right)^{-1} \quad \text{s. t.} \quad \sum_{\ell=1}^{L}\tau_\ell \leq P,$$

is easily shown to be reached at $\tau_\ell = \frac{P}{L}$, $\forall \ell$. This solution is independent of the realization of $\boldsymbol{q}$ and, therefore, it also minimizes Problem 5.24. According to this result, the general form of the optimum precoder in this case reads

$$\boldsymbol{T} = \sqrt{\frac{P}{L}}\boldsymbol{U},$$

where $\boldsymbol{U} \in \mathbb{C}^{N\times L}$ can be any matrix with orthonormal columns. Computing the average distortion for this choice of precoder we obtain

$$\epsilon = \frac{L}{P}\mathrm{E}_1\left(\frac{L}{P}\right)\exp\left(\frac{L}{P}\right). \tag{5.25}$$

Summing up, the optimum approach in this setting consists of allocating the same amount of power to all coefficients h_ℓ, $\ell = 1, \ldots, L$, and transmitting these coefficients over orthogonal precoders, which might or might not overlap in the frequency domain.[4] The minimum distortion does not depend on the dimensionality of the feedback channel but only on the transmit power and the dimension of the source outputs.

5.2.3.2 Flat Fading Forward Channel ($L = 1$)

In this case, the precoder reads $\boldsymbol{T} = \sqrt{P}\boldsymbol{u}$ with $\|\boldsymbol{u}\| \leq 1$. Define $w_m = \sum_{n=(m-1)N/M+1}^{mN/M} |u_n|^2$. We next demonstrate that any precoder such that $w_m = 1/M$, $m = 1, \ldots, M$, is optimum. Using the block fading assumption, Problem 5.23 simplifies to

$$\min_{\omega_1,\ldots,M} \mathrm{E}\left\{\frac{1}{1 + P\sum_{m=1}^{M}\omega_m \mathsf{z}_m}\right\} \quad \text{s. t.} \quad \sum_{m=1}^{M}\omega_m \leq 1,\ \omega_m \geq 0\ \forall m, \tag{5.26}$$

where z_m is the exponentially distributed fading coefficient corresponding to the mth block. This optimization problem is convex and, as a result, the KKT conditions are sufficient. The Lagrangian function of problem (5.26) reads

$$\begin{aligned} &L(\lambda, \mu_1, \ldots, \mu_M, \omega_1, \ldots, \omega_M) = \\ &= \int_{z_1,\ldots,z_M} \frac{1}{1 + P\sum_{m=1}^{M}\omega_m z_m} p(z_1, \ldots, z_M)dz_1 \cdots dz_M + \lambda\left(\sum_{m=1}^{M}\omega_m - 1\right) - \sum_{m=1}^{M}\mu_m\omega_m. \end{aligned}$$

[4] Also in the case that the entries of the source output h_ℓ, $\ell = 1, \ldots, L$, have unequal variance, it can be shown that optimum transmission is achieved by transmitting over orthogonal vectors. To prove this, first, the inequality $\mathrm{Tr}\{\boldsymbol{A}^{-1}\} \geq \mathrm{Tr}\{\mathrm{diag}\{\boldsymbol{A}\}^{-1}\}$ must be proved, where $\mathrm{diag}\{\boldsymbol{A}\}$ is the diagonal matrix built with the diagonal entries of $\boldsymbol{A}$. This inequality can be shown by noticing $\sum_i a_{ii}^{-1} \leq \sum_i \lambda_i^{-1}$, where $\{a_{ii}\}$ denote the diagonal entries in $\boldsymbol{A}$ and $\{\lambda_i\}$ denote the eigenvalues of this matrix. This last result follows directly from [61, Lemma 3.3.8]. Optimality of orthogonal precoding vectors follows by observing that, in particular,

$$\mathrm{Tr}\left\{\left(\boldsymbol{I} + q\boldsymbol{T}^{\mathrm{H}}\boldsymbol{T}\right)^{-1}\right\} \geq \mathrm{Tr}\left\{\left(\boldsymbol{I} + q\mathrm{diag}\left\{\boldsymbol{T}^{\mathrm{H}}\boldsymbol{T}\right\}\right)^{-1}\right\}.$$

Using the Leibniz rule to shift the derivative into the integral, the following optimality conditions can be computed,

$$-\int_{z_1,\ldots,z_M} \frac{Pz_j}{(1+P\sum_{m=1}^{M}\omega_m z_m)^2} p(z_1,\ldots,z_M)dz_1\cdots dz_M + \lambda - \mu_j = 0, \quad \forall j, \tag{5.27}$$

$$\lambda \geq 0, \quad \mu_m \geq 0 \ \forall m, \quad \lambda\left(\sum_{m=1}^{M}\omega_m - 1\right) = 0, \quad \mu_m\omega_m = 0 \ \forall m.$$

Choosing $\omega_m = 1/M, \ \forall m$, implies $\mu_m = 0, \ \forall m$. Now, noting that the integral in Eq. 5.27 is independent of index j, we observe that the condition represented by this equation is satisfied for all $j = 1,\ldots,M$ by appropriately choosing $\lambda \geq 0$.

Substituting $\omega_m = 1/M, \ \forall m$, in the objective function of Problem 5.26, for the minimum distortion, we obtain

$$\epsilon = \int_0^\infty \frac{p(\theta)}{1+P\theta}d\theta, \tag{5.28}$$

where $\theta = \frac{1}{M}\sum_{m=1}^{M} z_m$ is a chi-square random variable with $2M$ degrees of freedom and mean equal to 1. This integral can be expanded as a sum of exponential integral functions yielding

$$\epsilon = \frac{M^M \exp\left(\frac{M}{P}\right)}{(M-1)!P^M}\left(\sum_{m=0}^{M-2}\binom{M-1}{m}(-1)^m\alpha_{M-2-m}\left(\frac{M}{P}\right) + (-1)^{M-1}\mathrm{E}_1\left(\frac{M}{P}\right)\right), \tag{5.29}$$

where $\alpha_n(x) = \int_1^\infty t^n e^{-xt}dt$, $n = 0,1,2,\ldots$. Note that Eqs. 5.29 and 5.25 coincide if $M = 1$ and $L = 1$. Note also that distortion only depends of the transmit power and the diversity of the channel and is, as well as in the previous section, independent of the dimension of the feedback channel.

5.2.3.3 Moderately time-dispersive channels $LM \leq N$

In the previous section we have seen that, for $L = 1$, optimality is achieved by any precoder $\boldsymbol{T} = \sqrt{P}\boldsymbol{u}$ such that $w_m = \sum_{n=(m-1)N/M+1}^{mN/M} |u_n|^2 = 1/M$. In particular, it can be observed that in order to achieve optimality no more than M carriers are needed. If, now, $1 < L \leq N/M$, we can think of the following transmission strategy. Each coefficient h_ℓ can be transmitted over subcarriers $\ell + mN/M$, $m \in \{0,\ldots,M-1\}$, with power P_ℓ. That is, for each coefficient a set of M uncorrelated subcarriers is chosen for transmission, being the sets of subcarriers selected for transmission of different coefficients non-overlapping. Obviously, for a given fixed power allocation, $P_{1,\ldots,L}$, transmission is optimum as each coefficient is optimally transmitted.

In order to show optimality of the uniform power allocation we proceed as follows. Let $P_{1,\ldots,L}$ be any power allocation such that $\sum_\ell P_\ell = P$. The resulting minimum distortion for this allocation is given by

$$\epsilon = \frac{1}{L}\sum_{\ell=1}^{L}\psi(P_\ell) \quad \text{with} \quad \psi(P_\ell) = \mathrm{E}\left\{\frac{1}{1+P_\ell\theta}\right\}.$$

Noting that $\psi(P_\ell)$ is a convex function of P_ℓ, Jensens's inequality can be applied to obtain

$$\epsilon = \frac{1}{L}\sum_{\ell=0}^{L-1}\psi(P_\ell) \geq \psi(P/L).$$

Equality is achieved if the power allocation is uniform. In this case, the minimum distortion is in this case given by

$$\begin{aligned}\epsilon =& \frac{\exp\left(\frac{ML}{P}\right)}{(M-1)!}\left(\frac{ML}{P}\right)^M \times \\ & \times\left(\sum_{m=0}^{M-2}\binom{M-1}{m}(-1)^m\alpha_{M-2-m}\left(\frac{ML}{P}\right)+(-1)^{M-1}\mathrm{E}_1\left(\frac{ML}{P}\right)\right).\end{aligned} \tag{5.30}$$

Using Eq. 5.19, we note that for $L = N$ and $M = 1$ the distortion of the optimum analog approach is equal to the distortion of the upper bound given by Eq. 5.14, i.e., if the bandwidth expansion factor is 1 and the feedback channel is flat-fading, linear analog transmission performs optimally for all possible SNR values. This is the version for fading channels of the well-known result that linear analog transmission of a Gaussian source is optimum over the AWGN channel if the bandwidth expansion factor is 1 (cf. [57, 53]). Even though the model assumed here is very idealized, the following conclusions can be drawn that shall be useful for application in more realistic scenarios. First, in order to leverage diversity, each coefficient should be transmitted over the maximum possible number of statistically independent feedback channel coefficients, i.e., if the entries of matrix G represent the channel gains viewed on the different subcarriers of the feedback link, subcarriers used for transmission of a forward link channel coefficient should be separated by at least one coherence bandwidth. Second, transmission of a coefficient over different strongly correlated subchannels is likely to yield almost no gains in terms of performance. Note that in our model, for which subchannels in each block are fully correlated, ML subcarriers are enough to achieve optimality.

5.2.3.4 Asymptotical analysis

If $M = 1$, from Eqs. 5.15 and 5.29 we obtain

$$\epsilon(\mathrm{dB}) = -P(\mathrm{dB}) + O(\log(P(\mathrm{dB}))), \quad P \to \infty. \tag{5.31}$$

For $M > 1$, using the identity [1]

$$\alpha_n(x) = n!x^{-n-1}e^{-x}\left(1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!}\right),$$

it can be shown that distortion asymptotically behaves as

$$\epsilon = \frac{M}{(M-1)}\frac{L}{P} + o\left(\frac{1}{P}\right)$$

at high SNR values. That is, contrary to the theoretical upper bounds derived in Section 5.2.2, the optimum linear analog transmission scheme is not able to profit from the available diversity or the bandwidth expansion factor in terms of asymptotical distortion decay rate.

In the following we want to look at the asymptotic behavior of the distortion in the low SNR regime. Substituting $y = P/ML$ in Eq. 5.30, distortion can be written as

$$\epsilon = \frac{\exp(1/y)}{(M-1)!y^M}\left(\sum_{m=0}^{M-2}\binom{M-1}{m}(-1)^m\alpha_{M-2-m}(1/y) + (-1)^{M-1}\mathrm{E}_1(1/y)\right). \tag{5.32}$$

This expression can be expanded as a power series by using the expansions [1]

$$\exp(1/y)\alpha_n(1/y) = \frac{n!}{0!}y^{n+1} + \frac{n!}{1!}y^n + \frac{n!}{2!}y^{n-1} + \ldots + \frac{n!}{n!}y, \tag{5.33}$$

$$\exp(1/y)E_1(1/y) = 0!y - 1!y^2 + \ldots + (-1)^{n-1}(n-1)!y^n + \ldots. \tag{5.34}$$

Substituting Eqs. 5.33 and 5.34 in Eq. 5.32 we obtain

$$\epsilon = \frac{1}{(M-1)!y^M}\left((-1)^{M-1}\sum_{m=M}^{\infty}(-1)^{m-1}(m-1)!y^m\right).$$

Finally, shifting the quotient $\frac{1}{(M-1)!y^M}$ into the parenthesis and back substituting $y = P/ML$ the following expression results,

$$\epsilon = 1 - P/L + o(P),\ P \to 0.$$

That is, normalized distortion tends linearly to 1 as SNR $\to 0$ with slope $1/L$. This is exactly the same behavior as that of the theoretical upper bounds derived in Section 5.2.2. As a consequence, we conclude that linear analog transmission delivers optimum performance in the low SNR regime.

5.2.4 Delay-constrained digital transmission

Different from analog transmission approaches, digital transmission schemes employ a discrete and usually finite set of signals for transmission at the output of the encoder. Thus, the encoder is in this case a non-injective mapping that maps a set of values in the source space onto a single point or codeword in the signal space. As already discussed in Section 5.2.2, if no delay constraint is imposed, digital transmission achieves optimality. This is a consequence of the potential of digital transmission approaches of performing error-free transmission if delay is unbounded. Another consequence of error-free transmission is the separability of source and channel coding. Certainly, if no errors occur over the channel it does not matter how the mapping between source and channel codewords is made. As long as the mapping is bijective, it can be perfectly reversed at the receiver. That means that the source encoder needs not know the channel codebook and the channel encoder needs not know the source codebook. However, if delay is strictly limited, the probability of error is strictly larger than zero. In such case, it is convenient that channel codewords

being frequently mistaken for each other represent source values being close to each other in order to minimize distortion if transmission errors occur. Conversely, codewords that are unlikely to be mutually mistaken could represent very distant source values. That is, for delay limited digital transmission, the separation principle may lead to poor performance and, thus, care must be taken on how to map source values onto channel codewords. A theoretical framework for the analysis of digital systems with constrained delay and complexity is so far missing. An attempt to elaborate a general framework in [51] ended up with more questions than answers despite the simplifying assumption of a noiseless transmission channel. This lack of theoretical foundation has given rise to a heterogeneous landscape of approaches specifically tailored for particular settings and applications that are commonly referred to as joint source and channel coding schemes [59, 150].

Coming back to our model of Fig. 5.1, if digital transmission is considered, the encoder becomes a map from the source space onto a signal set $\mathcal{S} = \{\boldsymbol{s}_i \in \mathbb{C}^N | i = 1, \ldots, S\}$, i.e., for each output $\boldsymbol{h} \in \mathbb{C}^L$ of the source, the encoder chooses one of the S elements of the alphabet $\mathcal{S}$ for transmission. Based on the received signal, the decoder computes an estimate $\hat{\boldsymbol{h}} \in \mathbb{C}^L$ of the original source value $\boldsymbol{h}$. The goal is still the minimization of average MSE. However, now, rather than just having a transmit matrix to optimize over, optimization can be performed over the choice of map, size S of the constellation of transmit signals, and the choice of signals themselves subject to the transmit power constraint

$$\sum_{i=1}^{S} \|\boldsymbol{s}_i\|_2^2 P(\boldsymbol{s}_i) \leq P, \tag{5.35}$$

where $P(\boldsymbol{s}_i)$ is the probability that $\boldsymbol{s}_i$ is transmitted. A similar model has been extensively investigated in the literature considering a discrete memoryless channel (DMC) or binary symmetric channel (BSC) rather than an AWGN or fading channel [70, 47, 153, 152, 60]. This is equivalent to fixing the signal set in Fig. 5.1 and employing a hard decision detector as a first stage of the decoder. The AWGN channel has been considered in [125] where optimization of the encoder, decoder and signal set is performed assuming a linear decoder and a given constellation size. It turns out that, under the linearity assumption, average distortion achieved by the digital approach is bounded from below by that achieved by the linear analog approach. Performance of the analog approach is reached by the digital scheme as the constellation size approaches infinity. Therefore, recalling the asymptotic behavior of linear analog approaches, we conclude that under a linearity assumption at the receiver, the optimum delay constrained digital approach is incapable to profit from the bandwidth expansion factor in terms of distortion decay rate. In this respect, an interesting question is to know whether delay limited digital schemes may exhibit the asymptotic behavior of the theoretical upper bounds based on rate-distortion theory in case the linearity assumption is dropped. In [73] and [74] algorithms are proposed to optimize encoder, decoder and constellation for a given constellation size without any assumption on the structure of the decoder over AWGN and Rayleigh fading channels, respectively. The algorithms consist of the iterative alternating optimization of encoder and signal constellation and their execution relies on very expensive Monte Carlo methods. None of these works provide insights regarding fundamental performance of the resulting digital approaches. In the next sections, first, some details about the architecture of delay-constrained digital schemes are

discussed and the optimum decoder is derived. Then, a lower bound on the asymptotical decay rate of delay-constrained digital schemes is derived. Finally, two different paradigms for the design of the encoder are presented and discussed.

5.2.4.1 Arquitecture and optimum decoder

In the previous section, a memoryless encoder has implicitly been assumed in our delay-constrained digital model, i.e., the encoder is a map $\Phi : \mathbb{C}^L \to \mathcal{S}$ mapping the output of the source $\boldsymbol{h}[k]$ at instant k to one of the signals of the alphabet $\mathcal{S}$. Without loss of optimality, this map can be split in two blocks as illustrated in Fig. 5.2. A quantizer comprising S quantization cells Ω_i, $i = 1, \ldots, S$, which maps a source output onto the reproduction value $\boldsymbol{c}_i$ of the corresponding partition cell Ω_i, and a function $\phi : \{\boldsymbol{c}_i | i = 1, \ldots, S\} \to \mathcal{S}$ assigning one of the transmit signals to each of the reproduction values.

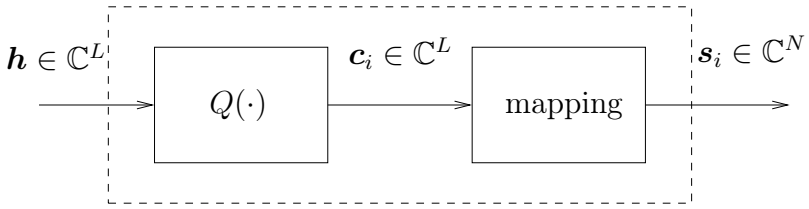

Figure 5.2: Encoder.

For fixed encoder, the decoder, which is assumed to be memoryless, takes the received signal $\boldsymbol{r}[k]$ at instant k and, using this observation, computes a minimum variance estimate of $\boldsymbol{h}[k]$ as follows,

$$\begin{aligned}
\hat{\boldsymbol{h}} &= \mathrm{E}\{\boldsymbol{h}|\boldsymbol{r}\} \\
&= \int \boldsymbol{h} p(\boldsymbol{h}|\boldsymbol{r}) d\boldsymbol{h} \\
&= \int \boldsymbol{h} \frac{p(\boldsymbol{r}|\boldsymbol{h}) p(\boldsymbol{h})}{p(\boldsymbol{r})} d\boldsymbol{h} \\
&= \frac{1}{p(\boldsymbol{r})} \int \boldsymbol{h} \sum_{i=1}^{S} p(\boldsymbol{r}|\boldsymbol{h}, \boldsymbol{s}_i) p(\boldsymbol{h}, \boldsymbol{s}_i) d\boldsymbol{h} \\
&= \frac{1}{p(\boldsymbol{r})} \int \boldsymbol{h} \sum_{i=1}^{S} p(\boldsymbol{r}|\boldsymbol{s}_i) P(\boldsymbol{s}_i|\boldsymbol{h}) p(\boldsymbol{h}) d\boldsymbol{h} \\
&= \frac{1}{p(\boldsymbol{r})} \sum_{i=1}^{S} p(\boldsymbol{r}|\boldsymbol{s}_i) \int_{\Omega_i} \boldsymbol{h} p(\boldsymbol{h}) d\boldsymbol{h} \qquad (5.36) \\
&= \frac{1}{p(\boldsymbol{r})} \sum_{i=1}^{S} p(\boldsymbol{r}|\boldsymbol{s}_i) P(\boldsymbol{s}_i) \boldsymbol{c}_i. \qquad (5.37)
\end{aligned}$$

Without loss of generality it has been assumed that the reproduction values $\boldsymbol{c}_{1,\ldots,S}$ of the

quantizer in Fig. 5.2 are the centroids of the quantization cells, which are defined as

$$\boldsymbol{c}_i = \int_{\Omega_i} \boldsymbol{h} p(\boldsymbol{h}|\boldsymbol{s}_i) d\boldsymbol{h}.$$

Eq. 5.36 follows by noting

$$P(\boldsymbol{s}_i|\boldsymbol{h}) = \begin{cases} 1 & \boldsymbol{h} \in \Omega_i \\ 0 & \boldsymbol{h} \notin \Omega_i \end{cases}$$

and Eq. 5.37 follows by noting $p(\boldsymbol{h}) = p(\boldsymbol{h}|\boldsymbol{s}_i)P(\boldsymbol{s}_i)$ for $\boldsymbol{h} \in \Omega_i$.

While in the context of digital transmission over noiseless channels, no gain can be achieved from inserting memory in the encoder or decoder if the source is memoryless [51, 72], it is unclear whether this result also holds if transmission takes place over a noise channel. That is, it is unclear that our assumption of memoryless encoder and decoder is without loss of optimality. If only the decoder is allowed to have memory, i.e., the decoder uses observations $\boldsymbol{r}[k-1], \boldsymbol{r}[k-2], \ldots$ in order to estimate $\boldsymbol{h}[k]$, while $\boldsymbol{r}[k]$ only depends on $\boldsymbol{h}[k]$, it is clear that an improvement with respect to an estimate exclusively based on $\boldsymbol{r}[k]$ is not possible as $\boldsymbol{r}[k-1], \boldsymbol{r}[k-2], \ldots$ are statistically independent with respect to $\boldsymbol{h}[k]$. Also in the case of a memoryless decoder and an encoder with memory, i.e., $\boldsymbol{h}[k]$ is estimated based on $\boldsymbol{r}[k]$ and $\boldsymbol{r}[k]$ corresponds to a transmit signal selected having into account $\boldsymbol{h}[k], \boldsymbol{h}[k-1], \boldsymbol{h}[k-2], \ldots$, no improvement can be expected. In this case $\boldsymbol{h}[k-1], \boldsymbol{h}[k-2], \ldots$ are nothing but randomizers of the map represented by the encoder, thus, acting as a source of noise from the point of view of the memoryless decoder. If both encoder and decoder have memory, the question becomes really interesting and challenging. In such case, observations $\boldsymbol{r}[k-1], \boldsymbol{r}[k-2], \ldots$ can help to better discern the signal transmitted by the encoder at time k, which also depends of $\boldsymbol{h}[k-1], \boldsymbol{h}[k-2], \ldots$ and this fact might help to improve the estimate of $\boldsymbol{h}[k]$. The cost of this is that at time k part of the resources are dedicated to carry information about $\boldsymbol{h}[k-1], \boldsymbol{h}[k-2], \ldots$, thereby deviating from the primary objective of reporting information about $\boldsymbol{h}[k]$.

5.2.4.2 Lower bound on asymptotic decay rate

In this section we derived a lower bound on the asymptotic decay rate of delay-constrained digital approaches by assuming suboptimum, though tractable, structures for the encoder and decoder. In the encoder, the quantizer is assumed to be an optimum MSE scalar quantizer, i.e., each real dimension is quantized independently using b bits. Denote the reproduction values by $\boldsymbol{q} = [\ q_1 \quad \cdots \quad q_{2L}\ ]^{\mathrm{T}}$ with $q_\ell \in \{c_j | j = 1, \ldots, 2^b\}$. Here, a real-valued representation of the reproduction values has been chosen for notational convenience. For each transmission a set of $S = 2^{2Lb}$ signals $\boldsymbol{s}_i \in \mathbb{C}^N$ is randomly generated. Each component of the signal vectors is independently drawn according to a circularly symmetric Gaussian distribution $\mathcal{CN}(0, P/N)$. The mapping block in Fig. 5.2 randomly maps the set of reproduction values onto set of randomly generated transmit signals. The decoder performs a maximum likelihood detection and reverses the random mapping in order to retrieve the transmitted quantizer output.[5]

[5]Note that for this scheme common randomness at encoder and decoder is a prerequisite.

Let $\hat{\boldsymbol{q}}$ be the reproduction value obtained at the receiver upon detection and mapping reversal being $\boldsymbol{q}$ the output of the quantizer. If no transmission errors occur $\hat{\boldsymbol{q}} = \boldsymbol{q}$ and distortion is entirely caused by the quantizer. At high resolution, i.e., high b, the distortion per complex dimension is well approximated by [55]

$$\epsilon_{\mathrm{ne}} = \frac{1}{2^{2b}12}\left\{\left(\int_{-\infty}^{\infty} p(h_{\mathrm{r}})^{1/3} dh_{\mathrm{r}}\right)^3 + \left(\int_{-\infty}^{\infty} p(h_{\mathrm{i}})^{1/3} dh_{\mathrm{i}}\right)^3\right\} = \frac{2\pi 3^{3/2}}{2^{2b}12}.$$

where h_{r}, h_{i} represent the real and imaginary parts of any of the entries of the source output. Consistently with the distribution of these entries, these variables are identically distributed as $\mathcal{N}(0, 1/2)$. If transmission errors occur, $\hat{\boldsymbol{q}} \neq \boldsymbol{q}$. Furthermore, due to the random mapping of reproduction values to signals at the encoder, $\hat{\boldsymbol{q}}$ is uniformly distributed among all reproduction values different from $\boldsymbol{q}$. That is, conditioned on the occurrence of a transmission error, all reproduction values different from $\boldsymbol{q}$ have the same probability to become the output of the decoder. Taking into account that distortion will in this case depend on the number of components of the original reproduction value that are erroneously detected, we can write

$$\epsilon = \epsilon_{\mathrm{ne}}(1 - P_{\mathrm{e}}) + \sum_{\ell=1}^{2L} P_{\mathrm{e}\ell}\epsilon_{\mathrm{e}\ell}. \tag{5.38}$$

There, $\epsilon_{\mathrm{e}\ell}$ is the average distortion per complex dimension conditioned on the erroneous detection of ℓ of the $2L$ components of the transmitted reproduction value. The probability that this happens can be computed as

$$P_{\mathrm{e}\ell} = \binom{2L}{\ell}\frac{(2^b - 1)^\ell}{2^{2bL} - 1}P_{\mathrm{e}},$$

where P_e is the probability of transmission error. Substituting this expression into Eq. 5.38, we obtain an expression of the distortion of the system as a function of the transmission error probability as follows,

$$\epsilon = \epsilon_{\mathrm{ne}}(1 - P_{\mathrm{e}}) + \bar{\epsilon}_{\mathrm{e}}P_{\mathrm{e}}, \tag{5.39}$$

where

$$\bar{\epsilon}_{\mathrm{e}} = \sum_{\ell=1}^{2L}\binom{2L}{\ell}\frac{(2^b - 1)^\ell}{2^{2bL} - 1}\epsilon_{\mathrm{e}\ell}$$

can be viewed as the average distortion per complex dimension conditioned on the occurrence of a transmission error.

In the following, the derivation of an upper bound for P_{e} follows along the lines of [66]. The only difference is that we allow codewords to violate the power constraint as long as this constraint is fulfilled in average. This somehow simplifies the analysis and the resulting expressions, and is consistent with common practice. Let $\boldsymbol{z} = [z_1, z_2, \cdots, z_M]$ represent the state of the block fading feedback channel at a particular time instant with $z_m \sim \mathcal{CN}(0, 1)$. $P_{\mathrm{e}} = \mathrm{E}\{P_e(\boldsymbol{z})\}$, where $P_e(\boldsymbol{z})$ is the error probability conditioned on the state $\boldsymbol{z}$. This probability can be shown to be upper bounded by

$$P_{\mathrm{e}}(\boldsymbol{z}) \leq \prod_{m=1}^{M} 2^{-NE_r(R|z_m)/M} \tag{5.40}$$

where $R = 2bL/N$ is the transmission rate and $E_r(R|z_m)$ is the random coding exponent corresponding to the mth chunk conditioned on the state z_m [66]. This exponent can be written as

$$E_r(R|z_m) = \max_{0\leq\rho\leq 1} \{E_0(\rho,\gamma|z_m) - \rho R\},$$

where

$$E_0(\rho,\gamma|z_m) = -\log_2 \int_{\mathbb{C}} \left(\int_{\mathbb{C}} \gamma(s) p(r|s,z_m)^{\frac{1}{1+\rho}} ds \right)^{1+\rho} dr, \tag{5.41}$$

and $\gamma(s)$ is the input distribution according to which transmit signals are generated [52]. Substituting $\gamma(s) = (\pi P/N)^{-1} \exp -N|s|^2/P$ and $p(y|s,z_m) = \pi^{-1} \exp -|y - z_m s|^2$ in Eq. 5.41 and computing the integrals we obtain

$$E_r(R|z_m) = \rho \left(\log_2 \left(1 + \frac{|z_m|^2 P/K}{1+\rho} \right) - R \right), \tag{5.42}$$

with $0 \leq \rho \leq 1$. If now Eq. 5.42 is substituted in Eq. 5.40 and the expectation is computed over the channel states, we get $P_\mathrm{e} \leq 2^{-M\bar{E}_r(R)}$, where

$$\bar{E}_r(R) = \max_{0\leq\rho\leq 1} \left\{ -\log_2 \left(a_\rho^{N\rho/M} \mathrm{e}^{a_\rho} \Gamma\left(1 - \rho N/M, a_\rho\right) \right) - \rho NR/M \right\}$$

and $a_\rho = (1+\rho)N/P$. Choosing $\rho = 1$ and using the identity $\Gamma(a,x) = x^a \mathrm{E}_{1-a}(x)$ and the inequalities $\mathrm{e}^x \mathrm{E}_n(x) \leq (x + n - 1)^{-1}$ and $\mathrm{e}^x \mathrm{E}_1(x) \leq \log_\mathrm{e}(1 + x^{-1})$, we can write $P_\mathrm{e} \leq 2^{-M(\bar{R}_0 - NR/M)}$ with

$$\bar{R}_0 = \begin{cases} \log_2 \left(1 + \frac{(N/M-1)P}{2N} \right), & N/M > 1 \\ \log_2 \left(\frac{P}{2N \log_\mathrm{e}(1+P/2N)} \right), & N/M = 1 \end{cases}. \tag{5.43}$$

This bound on the transmission error probability allows to upper bound Eq. 5.39 as

$$\epsilon \leq K(1 - P_\mathrm{e}) 2^{-NR/L} + \bar{\epsilon}_\mathrm{e} 2^{-M(\bar{R}_0 - NR/M)},$$

where $K = 2\pi 3^{3/2}/12$. Now, choosing $b = \left\lfloor \frac{M\bar{R}_0}{2(L+1)} \right\rfloor$ and noting $(1 - P_\mathrm{e}) \leq 1$ the following upper bound results,

$$\epsilon \leq (K + \bar{\epsilon}_\mathrm{e}) 2^{-2\left\lfloor \frac{M\bar{R}_0}{2(L+1)} \right\rfloor}.$$

Assume that the average distortion conditioned on transmission errors $\bar{\epsilon}_\mathrm{e}$ is bounded. In that case, considering Eqs. 5.43, it is easily shown that for the derived upper bound on distortion

$$\mathrm{DDR} = \frac{M}{L+1}$$

holds (cf. Eq. 5.16). That is, the distortion decay rate of the optimum delay constrained digital approach is lower bounded by $M/(L+1)$. This shows that, different from linear analog approaches, delay-constrained digital approaches have the potential to profit from the diversity degree available in the feedback link in terms of distortion decay rate.

The boundedness assumption on $\bar{\epsilon}_\mathrm{e}$ is key for the validity of this result. $\bar{\epsilon}_\mathrm{e}$ is bounded if all $\epsilon_{\mathrm{e}\ell}$, $\ell = 1, \ldots, 2L$, are bounded. Obviously, among these average distortion values, the

largest is ϵ_{e2L}. Thus, it is enough to show that ϵ_{e2L} is bounded. Let $\boldsymbol{h} = [\ h_1 \quad \dots \quad h_{2L}\ ]^{\mathrm{T}}$ be the source output, which for notational convenience we represent as a vector of $2L$ real dimensions. Let I_{n_ℓ} be the quantization interval corresponding to the ℓth component of the source output and c_{n_ℓ} the corresponding reproduction value. If this source output is transmitted the average distortion per complex dimension at the receiver conditioned on the occurrence of $2L$ errors is given by

$$\epsilon_{e2L}(\boldsymbol{h}) = \frac{1}{L}\sum_{\ell=1}^{2L}\frac{1}{2^b-1}\sum_{\substack{j=1\\ j\neq n_\ell}}^{2^b}(h_\ell - c_j)^2, \tag{5.44}$$

where we have taken into account the fact that, conditioned on a detection error, all reproduction values apart from transmitted one are equally likely. Averaging this expression over all possible outputs of the source we obtain

$$\epsilon_{e2L} = \frac{2}{2^b-1}\sum_{n=1}^{2^b}\int_{I_n}\sum_{\substack{j=1\\ j\neq n}}^{2^b}(h-c_j)^2 p(h)dh \tag{5.45}$$

where the fact has been used that all $2L$ terms in Eq. 5.44 are identically distributed and, as a result, index ℓ has been dropped. Expanding the square in Eq. 5.45 and after some simple manipulations we obtain

$$\epsilon_{e2L} = 1 + \frac{2}{2^b-1}\sum_{j=1}^{2^b}c_j^2 + \frac{2}{2^b-1}\left(2\sum_{j=1}^{2^b}\int_{I_j}c_j h p(h)dh - \sum_{j=1}^{2^b}\int_{I_j}c_j^2 p(h)dh\right).$$

In order to prove boundedness it is enough to prove convergence of ϵ_{e2L} as $b\to\infty$. This is a consequence of the well known mathematical result that every convergent sequence is bounded. As $b\to\infty$, $c_j \to h$ in the integrals of the third term. As a result,

$$\sum_{j=1}^{2^b}\int_{I_j}c_j h p(h)dh \to 1/2,$$

$$\sum_{j=1}^{2^b}\int_{I_j}c_j^2 p(h)dh \to 1/2,$$

and the third term converges to zero. As for the second term, boundedness is proved as follows. First, we can write

$$\frac{1}{2^b}\sum_{j=1}^{2^b}c_j^2 = \frac{1}{2^b}\sum_{j=1}^{2^b}\frac{c_j^2}{\Delta_j}\Delta_j \tag{5.46}$$

where Δ_j is the length of I_j. If $b\to\infty$, in the numerator, we substitute Δ_j by dh. In the denominator, we apply the approximation [55]

$$\Delta_j \approx \frac{1}{2^b\lambda(c_j)}$$

where $\lambda(h)$ is the point density function of the optimum scalar quantizer at high resolution[6].

[6] $\lambda(h) = \frac{p(h)^{1/3}}{\int p(h)^{1/3}dh}$

Proceeding this way we observe[7]

$$\frac{1}{2^b}\sum_{j=1}^{2^b} h_j^2 \to \int h^2 \lambda(h) dh < \infty, \quad b \to \infty.$$

5.2.4.3 Encoder design paradigms

To find the optimum encoder for delay-limited digital transmission is by all means a very hard problem. Nevertheless, meaningful design solutions can be found that yield acceptable performance. In particular, given a fixed quantizer, in this section we discuss two different design paradigms for the mapping block of Fig. 5.2. The first paradigm will be referred to as topological and aims at mapping reproduction values onto transmit signals so that the neighborhood relations are preserved over the map. The rationale of this approach is simple. Transmit signals which are close to each other are likely to be mutually mistaken. However, since they represent reproduction values that are close to each other, distortion resulting from mistaking neighboring signals remains moderate. A second paradigm, which we call non-topological, aims at maximizing the distance between any two transmit signals in the image of the map while meeting the transmit power constraint. In this way, transmit signals can be more clearly distinguish at the receiver, thus, reducing distortion arising from the noisy channel. Ideally, the combination of both paradigms should deliver a design close to optimum where communication is reliable and mistaking neighboring signals has a mild impact. Unfortunately, maximizing the minimum distance between any two points of the transmit signal set in a space of dimension $N > L$ seems to inevitable lead to an increase in the number of neighboring points to any other given, thereby destroying the neighborhood relations of the original set of reproduction values. As stated in [102], any system which attempts to use the capacities of an increase in dimensions to the full possible extent seems to be bound to suffer from the threshold effect.[8]

The difference between both mappings is illustrated in Fig. 5.3. There, every quantizer reproduction value is represented by a different marker. A quantizer point is mapped to the point in the signal space represented by the same marker. Points in the signal space have two complex dimensions. Let us pay attention to the point represented by the dot on the upper left corner of the quantizer. In the image obtained through the topological mapping the points represented by the circle and the plus sign are still the ones with minimum distance to the dot. As in the original domain, the distance to any other point is strictly larger. Being d the minimum distance between two points in one of the complex dimensions, the distance between the dot and any of its two neighbors is given by $\sqrt{2}d$. If the non topological mapping is applied, the minimum distance between the dot and any

[7] In the derivation of this result there is a shortcoming that has been here ignored for clarity of exposition. In Eq. 5.46 two of the intervals have infinite length. In order to overcome this technical difficulty a truncated Gaussian distribution can be used, for which the derivation is technically accurate. Noting convergence of the truncated Gaussian distribution towards a Gaussian distribution when the truncation interval tends to infinity the result follows.

[8] The threshold effect is the typical abrupt performance degradation of coded transmission when the SNRs decreases below a certain value. This degradation is due to the high number of neighboring points in packings of signals with good distance properties [36].

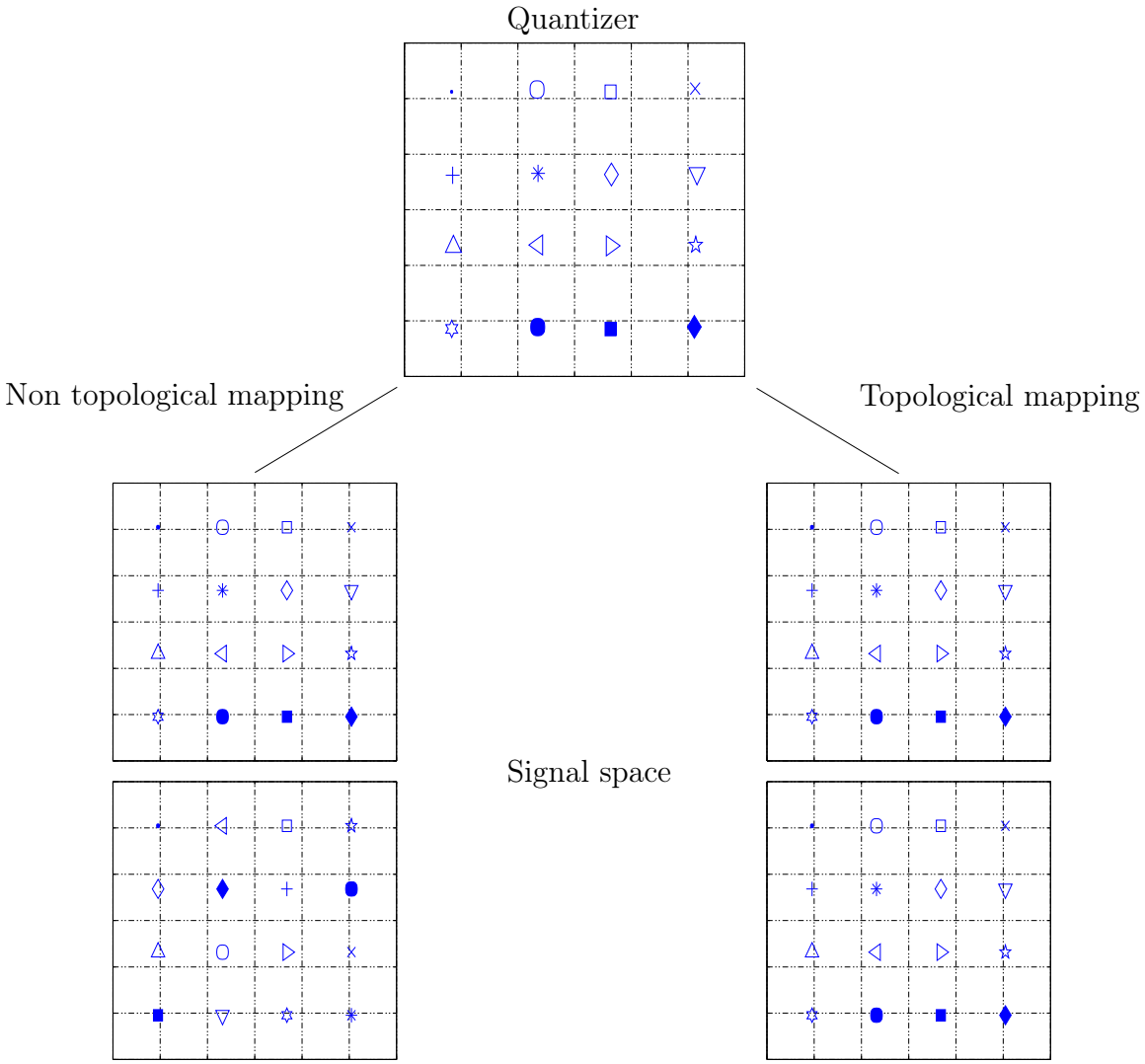

Figure 5.3: Topological and non topological mappings for $L = 1$, $S = 16$ and $N = 2$.

other point in the image of this mapping increases to $\sqrt{6}d$. However, instead of two, now the number of neighbors increases to four, viz., plus sign, circle, diamond and triangle. That is, the neighborhood relations in the domain are not preserved in the image. There are even points that become closer to a certain point in the range relative to other that were closer to that point in the domain. This is, for instance, the case of the asterisk and the diamond with respect to the dot.

While non-topological mappings can be obtained by using standard coded modulation schemes in order to generate signal sets with good distance properties, a simple way of obtaining topological mappings consists of linearly transforming the reproduction values employing a scaled matrix with orthonormal columns, i.e.,

$$\boldsymbol{s}_j = \alpha \boldsymbol{U} \boldsymbol{c}_j, \quad j = 1, \ldots, S,$$

where $\boldsymbol{U} \in \mathbb{C}^{N \times L}$ and $\alpha = P / \sum_j \|\boldsymbol{c}_j\|_2^2 P(\boldsymbol{c}_j)$ is a scaling factor that guarantees fulfillment

of the transmit power constraint. This transformation preserves the distance relations of the points in the domain. However, in the following we shall see that the maximum DDR achievable by this mapping is 1.

Recall Eq. 5.37 and note that

$$p(\boldsymbol{r}) = \sum_{i=1}^{S} p(\boldsymbol{r}|\boldsymbol{s}_i)P(\boldsymbol{s}_i). \tag{5.47}$$

Substituting Eq. 5.47 in Eq. 5.37 we obtain

$$\hat{\boldsymbol{h}}(\boldsymbol{r}) = \frac{\sum_{i=1}^{S} p(\boldsymbol{r}|\boldsymbol{s}_i)P(\boldsymbol{s}_i)\boldsymbol{c}_j}{\sum_{i=1}^{S} p(\boldsymbol{r}|\boldsymbol{s}_i)P(\boldsymbol{s}_i)}. \tag{5.48}$$

Further,

$$p(\boldsymbol{r}|\boldsymbol{s}_i) = \frac{1}{\pi^N} \exp -\|\boldsymbol{r} - \boldsymbol{G}\boldsymbol{s}_i\|_2^2,$$

and the norm of the exponent can be expressed as

$$\begin{aligned} \|\boldsymbol{r} - \boldsymbol{G}\boldsymbol{s}_i\|_2^2 &= \|\boldsymbol{r}\|_2^2 + \alpha^2 \boldsymbol{c}_i^{\mathrm{H}}\boldsymbol{U}^{\mathrm{H}}\boldsymbol{G}^{\mathrm{H}}\boldsymbol{G}\boldsymbol{U}\boldsymbol{c}_i - \alpha\boldsymbol{c}_i^{\mathrm{H}}\boldsymbol{U}^{\mathrm{H}}\boldsymbol{G}^{\mathrm{H}}\boldsymbol{r} - \alpha\boldsymbol{r}^{\mathrm{H}}\boldsymbol{G}\boldsymbol{U}\boldsymbol{c}_i \\ &= \|\tilde{\boldsymbol{r}} - \alpha\boldsymbol{\Lambda}\boldsymbol{V}\boldsymbol{c}_i\|_2^2 + \|\boldsymbol{r}\|_2^2 - \|\tilde{\boldsymbol{r}}\|_2^2 \\ &= \|\tilde{\boldsymbol{r}} - \alpha\boldsymbol{M}\boldsymbol{c}_i\|_2^2 + f(\boldsymbol{r}). \end{aligned} \tag{5.49}$$

There, $\boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{V}$ is the singular value decomposition of $\boldsymbol{G}\boldsymbol{U}$, with $\boldsymbol{\Lambda} \in \mathbb{R}^{L\times L}$, $\tilde{\boldsymbol{r}} = \boldsymbol{W}\boldsymbol{r} \in \mathbb{C}^L$ and $\boldsymbol{M} = \boldsymbol{\Lambda}\boldsymbol{V} \in \mathbb{C}^{L\times L}$. Finally, substituting Eq. 5.49 in Eq. 5.48 we get

$$\hat{\boldsymbol{h}}(\boldsymbol{r}) = \hat{\boldsymbol{h}}(\tilde{\boldsymbol{r}}) = \frac{\sum_{i=1}^{S} p(\tilde{\boldsymbol{r}}|\boldsymbol{c}_i)P(\boldsymbol{c}_i)\boldsymbol{c}_i}{\sum_{i=1}^{S} p(\tilde{\boldsymbol{r}}|\boldsymbol{c}_i)P(\boldsymbol{c}_i)} \tag{5.50}$$

where

$$p(\tilde{\boldsymbol{r}}|\boldsymbol{c}_i) = \frac{1}{\pi^L} \exp -\|\tilde{\boldsymbol{r}} - \alpha\boldsymbol{M}\boldsymbol{c}_i\|_2^2.$$

According to Eq. 5.50, $\tilde{\boldsymbol{r}}$ is a sufficient statistic of vector $\boldsymbol{r}$ for estimation of the source $\boldsymbol{h}$. In addition, $\tilde{\boldsymbol{r}}$ may be viewed as the received signal of an equivalent system with channel matrix $\boldsymbol{M}$. For a particular realization of $\boldsymbol{M}$ the capacity of this system is given by

$$C(\boldsymbol{M}) = \log_2\left(\left|\boldsymbol{I}_L + \boldsymbol{\Lambda}^2 P/L\right|\right).$$

Equating this expression and Eq. 5.8 and solving for ϵ we obtain the minimum achievable distortion over that particular channel realization as

$$\epsilon(\boldsymbol{M}) = \frac{1}{\left|\boldsymbol{I}_L + \boldsymbol{\Lambda}^2 P/L\right|^{1/L}}.$$

If $P > 1$,

$$\epsilon(\boldsymbol{M}) \geq \frac{1}{P\left|\boldsymbol{I}_L + \boldsymbol{\Lambda}^2/L\right|^{1/L}}.$$

From this expression we observe that the distortion decay rate for a particular channel realization is upper bounded by 1. As this upper bound is independent of the channel realization, we conclude that 1 is also an upper bound for the decay rate of the average distortion over all possible realizations of matrix M.

5.2.5 Numerical results

In order to illustrate the difference between topological and non-topological mappings we consider a simple setting with $L = 1$, $N = 4$ and $M = 1$. The source is coded with a scalar MMSE quantizer. The topological mapping is a linear map defined by $\boldsymbol{u} = [1/\sqrt{4}, \ldots, 1/\sqrt{4}]^{\mathrm{T}}$ and a scalar factor α that guarantees fulfillment of the power constraint. Simulations have been carried out with $b = 1, 2, 3, 4, 5, 6$ resolution bits per real dimension. The non-topological mapping uses rate $1/2$ block codes in order to code the binary labels corresponding to the quantizer reproduction values. Then, the resulting coded bits are segmented and mapped onto signal points of QAM constellations of suitable size. For $b = 1$, a $(4, 2)_2$ block code[9] with generator matrix

$$\begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

has been employed. For $b = 2$, the $(7, 4)_3$ Hamming code has been extended to a $(8, 4)_4$ code by adding an additional parity check bit to every code word [48]. For $b = 3, 4, 5$, block codes $(12, 6)_4$, $(16, 8)_4$ and $(20, 10)_6$ have been chosen with generator matrices

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \end{pmatrix},$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix},$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix},$$

[9] A code word of a $(n, k)_d$ block code has k information bits and is n bits long. The minimum Hamming distance between any two code words is given by d.

respectively. For $b = 6$ the Golay code $(23, 12)_7$ has been extended to a $(24, 12)_8$ by adding a parity check bit to every code word [48]. At the receiver an optimum MMSE estimator has been applied.

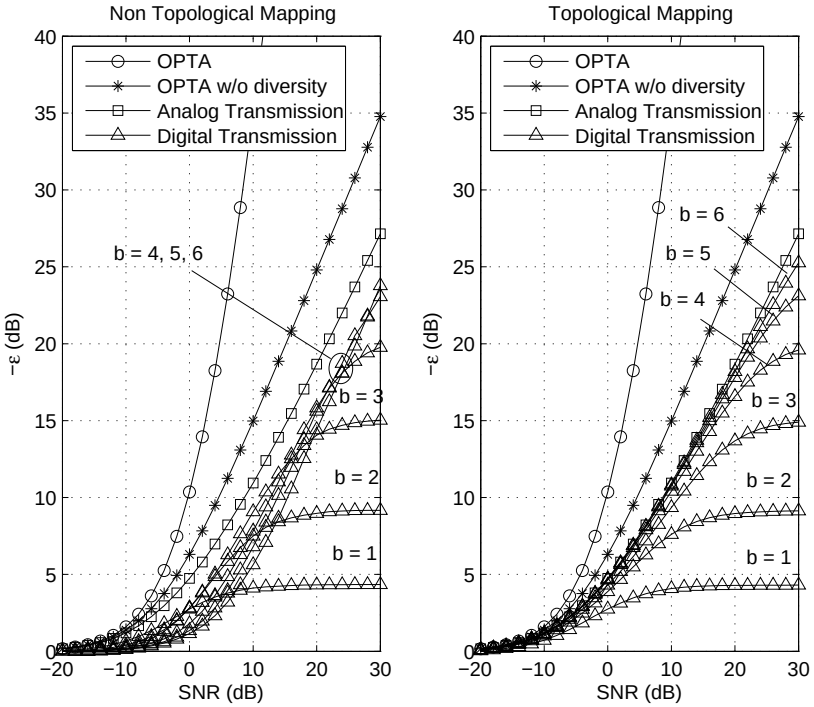

Figure 5.4: Performance of delay-constrained digital transmission. $L = 1$, $N = 4$, $M = 1$.

In Fig. 5.4 distortion curves are shown for both topological and non-topological mappings. The theoretical upperbounds derived in Section 5.2.2 and the curve corresponding to the linear analog scheme are also plotted. In the x-axis, SNR $= P/N$. As indicated by our analysis of Section 5.2.2 the asymptotic slope of the curve corresponding to OPTA is proportional to the bandwidth expansion factor, which is 4 in this case. By contrast, the slope of the curve corresponding to the theoretical upper bound with limited diversity is just 1. The slope of the analog approach is less than 1 (cf. Eq. 5.31) but tends to 1 asymptotically. In the low SNR regime, the curve of the analog scheme converges to the OPTA curves. For this setting, delay constrained digital approaches are in all cases outperformed by the analog transmission scheme. Performance of the topological approach is tightly upper bounded by performance of the analog scheme. At least in this setting, more bits lead to a uniform performance improvement of the topological scheme for all SNR values. This is not the case for the non-topological mapping, where curves corresponding to different resolutions exhibit crossover points. The non-topological mapping is clearly outperformed by the topological mapping. An explanation for this poor performance can be found in the threshold effect and the fact that for a given transmission rate good performance obtained with good channel realizations is gambled away by disastrous performance obtained with poor channel realizations.

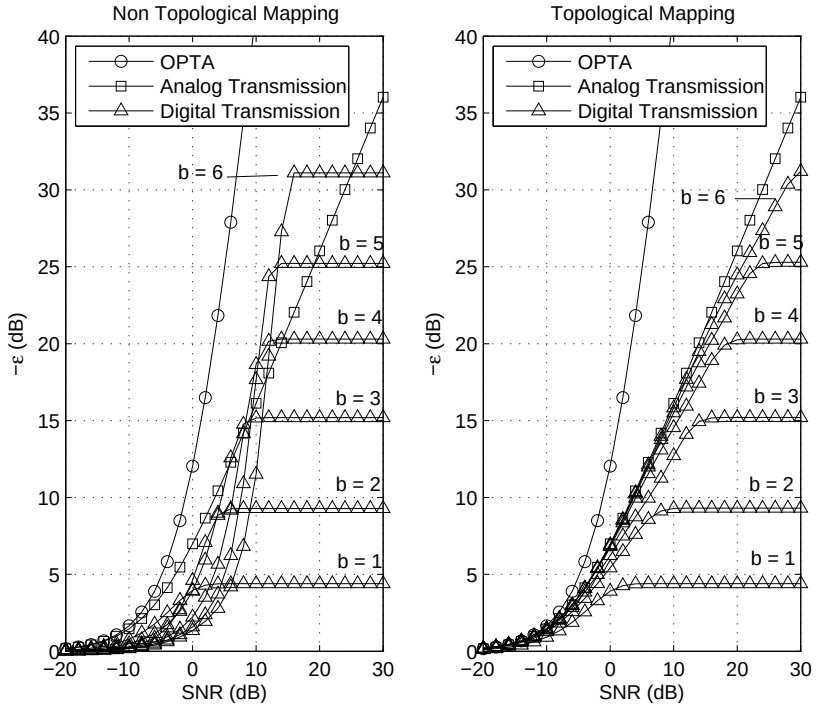

Figure 5.5: Performance of delay constrained digital transmission over an AWGN feedback channel. $L = 1$, $N = 4$.

For comparison purposes, Fig. 5.5 shows performance for the same setting when transmission takes place over an AWGN channel, i.e., the coefficients on the diagonal of $\boldsymbol{G}$ are fixed and equal to 1. As the channel does not fade, OPTA is the only applicable upper bound, which grows with slope 4. The threshold effect of the non-topological mapping becomes now very visible with performance breaking down abruptly within a few dBs. The topological mapping, by contrast, shows gracefully degradation, thereby outperforming the non-topological mapping at low SNR values. However, performance of the non-topological mapping seems to improve with increasing SNR as fast as OPTA, i.e., non-topological mappings are able to benefit from the bandwidth expansion factor, whereas the decay rate of the topological scheme is upper bounded by 1.

Fig. 5.6 shows simulation results for a setting with $L = 2$ and $N = 16$. The feedback channel is considered to be a time-dispersive channel with two taps and flat power delay profile. The coefficients of matrix G are the channel gains resulting from employing an OFDM transmission scheme over this channel with $N = 16$ subcarriers. As expected, the behavior of the simulated curves matches the analytical results obtained based on a block-fading model with $M = 2$. The OPTA curve grows with a slope that is between 4 and 8. By contrast, the asymptotic slope of the OPTA upper bound with limited diversity is just 2. The linear analog scheme has been applied by transmitting each coefficient over two subchannels being one coherence-bandwidth apart from each other. The resulting slope is equal to 1. At low SNR, the curve corresponding to the analog scheme converges to

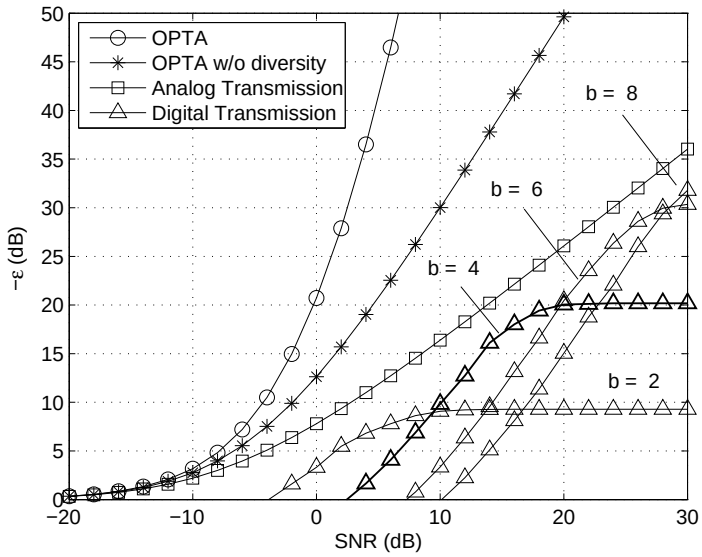

Figure 5.6: Performance over a fading feedback link with $L = 2$, $N = 16$, $M = 2$.

the upper bounds. For the delay-constrained digital transmission curves, a scalar MMSE quantizer has been considered with b resolution bits. Even in this simple setting, use of an optimum estimator at the receiver becomes extraordinarily complex for resolutions as low as $b = 4$ bits per real dimension. For this reason, a suboptimum decoder structure has been considered that first performs detection of transmitted bits and then maps the detected bits to reproduction values. The mapping at the encoder is performed according to a non-topological paradigm. The binary labels provided by the quantizer are encoded by using two convolutional codes concatenated in parallel as described in [6]. The resulting bits are randomly interleaved and mapped to points of QAM constellations of suitable sizes according to a bit interleaved coded modulation scheme [25]. Detection in the first stage of the decoder is performed iteratively as described in [6]. Due to the suboptimum decoder the curves disappear below the x-axis at low SNR values, i.e., distortion can be larger than 1. Nonetheless, increase in performance parallels that of the tighter upper bound in the high SNR regime. Increasing the number of taps in the feedback channel to $M = 4$ and keeping all other parameters, the curves plotted in Fig. 5.7 result. OPTA does not change as it does not depend on the diversity degree of the channel (cf. Eq. 5.10). Now, the curve corresponding to OPTA without temporal diversity exhibits an asymptotical slope equal to 4. The delay-constrained digital curves seem to parallel this growth before they reach saturation. However, due to the large offset between this curves and the tighter upper bound, the analog scheme exhibits better performance for almost all simulated SNR values. Note that some improvement in performance of the delay-constrained digital schemes can be attained if the scalar quantizer is replaced by a vector quantizer at the cost of additional complexity. However, due to the fact that the source outputs are statistically independent, only a modest gain of around 2 dBs is expected in this case [76].

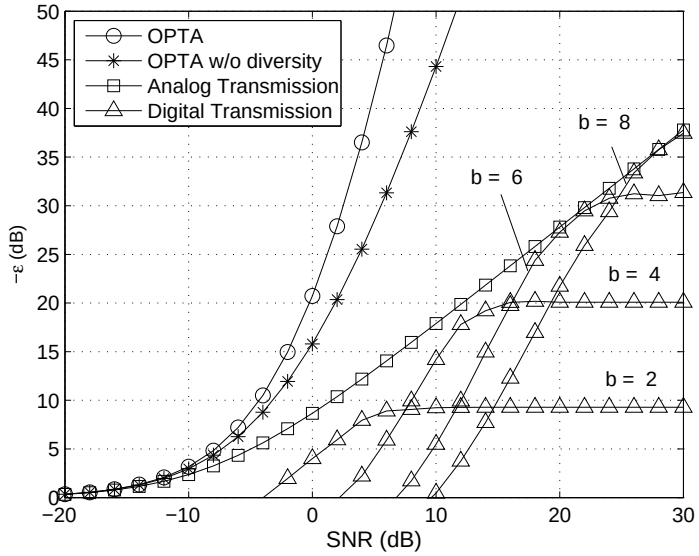

Figure 5.7: Performance over a fading feedback link with $L = 2$, $N = 16$, $M = 4$.

5.3 Extension to feedback channels with multiple antennas

In this section we extend some of the results obtained in the previous section to feedback channels with multiple antennas. The feeback link model remains as depicted in Fig. 5.1. The channel matrix $\boldsymbol{G}$ is no longer a diagonal matrix but a block diagonal matrix. The dimensions of the blocks on the diagonal are determined by the number of antennas at both ends of the feedback link. The dimensions of all other signals in the model are modified accordingly. In the following we shall study a feedback link which has one single antenna at the transmitter side and more than one antenna at the receiver side. The corresponding forward link may consist of a transmitter having multiple antennas and single-antenna receivers. The general MIMO setting with multiple antennas at both ends and the MISO setting with multiple antennas at the transmit end of the feedback link and a single-antenna receiver will not be treated. The former is hardly tractable requiring perhaps more sophisticated mathematical tools. The latter, though tractable, corresponds to a forward link with one transmit antenna and several receive antennas and, therefore, lacks practical interest. For the SIMO feedback link that we will investigate here, the channel matrix becomes an $Nt \times N$ block diagonal matrix with blocks $\boldsymbol{g}_n \in \mathbb{C}^{t\times 1}$, $n = 1, \ldots, N$. In order to denote the number of antennas at the receive end of the feedback link we have chosen t. This is consistent with notation used in the other chapters of this work as here we shall assume that this variable also represents the number of antennas at the transmitter of the corresponding forward link. The source outputs are now assumed to be vectors of dimension Lt and uncorrelated zero-mean circularly symmetric Gaussian distributed entries with unit variance. As in the initial feedback link model, source outputs at different

time instants are uncorrelated. Here, we will consider that a source output represents the channel state of a time-dispersive forward channel with flat power delay profile, L taps, t transmit antennas and no spatial correlation.

5.3.1 Theoretical upper bounds

The rate-distortion function of the source is now given by

$$R(\epsilon) = \begin{cases} Lt \log\left(\frac{1}{\epsilon}\right) & \text{if} \quad \epsilon < 1 \\ 0 & \text{if} \quad \epsilon \geq 1 \end{cases}. \tag{5.51}$$

The capacity of the feedback channel can be computed as [105]

$$C = \sum_{n=1}^{N} \mathrm{E}\left\{\log\left(1 + \frac{\|\boldsymbol{g}_n\|_2^2 P}{N}\right)\right\} = \frac{N}{\log_{\mathrm{e}} 2} \exp\left(\frac{N}{P}\right) \sum_{i=1}^{t} \mathrm{E}_i\left(\frac{N}{P}\right). \tag{5.52}$$

Equating Eqs. 5.51 and 5.52 and solving for ϵ we obtain the OPTA as

$$\epsilon_{\mathrm{OPTA}} = \exp\left(-\frac{N}{Lt} \exp\left(\frac{N}{P}\right) \sum_{i=1}^{t} \mathrm{E}_i\left(\frac{N}{P}\right)\right). \tag{5.53}$$

In the high SNR regime, using Eqs. 5.15 and 5.20, it can be easily shown that

$$\epsilon_{\mathrm{OPTA}}(\mathrm{dB}) = -\eta \frac{N}{Lt} P(\mathrm{dB}) + O(1), \quad P \to \infty,$$

with $1/2 \leq \eta \leq 1$. In the low SNR regime, substituting the bounds given in Eqs. 5.15 and 5.20 in Eq. 5.53 and computing a linear approximation of the resulting expressions around $P = 0$, we obtain

$$\epsilon_{\mathrm{OPTA}} = 1 - P/L + o(P). \tag{5.54}$$

We note that in the high SNR regime, performance degrades as the number of receive antennas increases, i.e., the advantage due to an increase in number of antennas in the feedback link does not compensate for the increase in the number of channel coefficients that must be fed back. In the low SNR regime, by contrast, performance is independent of the number of receive antennas.

As explained in Section 5.2.2.2, this bound can be tightened by limiting the amount of diversity to that available in just one use of the feedback link. The resulting OPTA without time diversity is computed by first considering the minimum distortion achievable for a particular realization of the feedback channel,

$$\epsilon(\boldsymbol{g}_{1,\ldots,N}) = \frac{1}{\prod_{n=1}^{N} \left(1 + \|\boldsymbol{g}_n\|_2^2 P/N\right)^{\frac{1}{Lt}}}.$$

and then averaging this expression over all possible channel realizations. Using the block fading assumption, the average distortion can be written as

$$\epsilon_{\mathrm{OPTA-LD}} = \left(\mathrm{E}\left\{\frac{1}{(1 + \boldsymbol{q}P/N)^{\frac{N}{MLt}}}\right\}\right)^{M}, \tag{5.55}$$

where $\boldsymbol{q}$ has the same distribution as $\|\boldsymbol{g}_n\|_2^2$, $\forall n$. Unfortunately, analytical computation of this expectation is very difficult, if possible at all. However, in the following, two lower bounds on distortion will be used that reveal the different behavior of this bound as compared to OPTA. In order to derive the first of these bounds, we note that $\boldsymbol{q}$ can be viewed as a sum of t statistically independent, exponentially distributed random variables. Let $\boldsymbol{q}_i$, $i = 1, \ldots, t$, denote each of these variables. We can write

$$\epsilon_{\text{OPTA-LD}} \geq \left(\mathrm{E}\left\{ \frac{1}{\prod_{i=1}^{t} (1 + \boldsymbol{q}_i P/N)^{\frac{N}{MLt}}} \right\} \right)^M = \left(\mathrm{E}\left\{ \frac{1}{(1 + \boldsymbol{z} P/N)^{\frac{N}{MLt}}} \right\} \right)^{Mt},$$

where $\boldsymbol{z}$ is exponentially distributed with mean equal to 1. Computation of the expected value in the last expression can be done as described in Section 5.2.2.2. Doing so, we obtain

$$\epsilon_{\text{OPTA-LD}}^{\frac{1}{Mt}} \geq \left(\frac{N}{P}\right)^{\frac{N}{MLt}} \exp\left(\frac{N}{P}\right) \Gamma\left(-\frac{N}{MLt} + 1, \frac{N}{P}\right).$$

In the high SNR regime, an analysis of this expression along the lines of that carried out in Section 5.2.2.3 yields

$$\mathrm{DDR} \leq \begin{cases} \frac{N}{L}, & \frac{N}{MLt} \leq 1 \\ Mt, & \frac{N}{MLt} \geq 1 \end{cases}.$$

In order to derive the second bound, we note that the argument of the expectation operator in Eq. 5.55 is a convex function of q. Hence, applying Jensen's inequality we can write

$$\epsilon_{\text{OPTA-LD}} \geq \left(\frac{1}{(1 + \mathrm{E}\{\boldsymbol{q}\} P/N)^{\frac{N}{MLt}}} \right)^M = \left(\frac{1}{(1 + tP/N)^{\frac{N}{MLt}}} \right)^M.$$

Using this bound, we can easily verify that $\mathrm{DDR} \leq \frac{N}{Lt}$ holds. If we select the most restrictive of these two bounds for each choice of parameters, the following bound results,

$$\mathrm{DDR} \leq \begin{cases} \frac{N}{Lt}, & \frac{N}{Lt} \leq Mt \\ Mt, & \frac{N}{Lt} \geq Mt \end{cases}.$$

That is, the asymptotic distortion decay rate is limited by either the bandwidth expansion factor $\frac{N}{Lt}$ or the diversity degree Mt of the feedback channel. This bound also suggests an interesting trade-off on the number of antennas. If t grows, diversity increases but the bandwidth expansion factor decreases. Conversely, if t decreases, the bandwidth expansion factor increases but performance limitation may be due to diversity, which decreases in this case. Obviously, the least restrictive condition is achieved if bandwidth expansion factor and diversity degree are both equal, i.e., $t = \sqrt{N/LM}$.

5.3.2 Analog transmission

Assuming $LMt \leq N$, the results obtained in Section 5.2.3 can easily be extended to the multiple receive antennas setting considered in this section. In order to derive the optimum

transmission strategy, we can proceed according to the steps followed in Section 5.2.3.2. Using the block fading model and under the assumption that only one channel coefficient is transmitted, Problem 5.23 can be written as Problem 5.26, where, now, the variables z_m are chi-square distributed with $2t$ degrees of freedom and mean value t. Noting that in the subsequent derivation of the optimum transmission strategy the specific distribution of z_m is not used, we conclude that choosing any precoder such as $w_m = 1/M$, $m = 1, \ldots, M$, is optimum also in this setting. Now, if transmission of the Lt coefficients of the source output is considered and $LMt \leq N$, the reasoning of Section 5.2.3.3 can be applied in order to find that optimality is achieved if coefficients are transmitted over disjoint sets of M uncorrelated subchannels with a uniform power distribution. In order to compute the distortion incurred by the optimum approach, Eq. 5.28 can be used replacing P by P/Lt and noting that θ is now a chi-square random variable with $2Mt$ and mean value t. Doing so, we obtain

$$\epsilon = \frac{\exp\left(\frac{MLt}{P}\right)}{(Mt-1)!}\left(\frac{MLt}{P}\right)^{Mt} \times \\ \times\left(\sum_{m=0}^{Mt-2}\binom{Mt-1}{m}(-1)^m\alpha_{Mt-2-m}\left(\frac{MLt}{P}\right)+(-1)^{Mt-1}\mathrm{E}_1\left(\frac{MLt}{P}\right)\right).$$

To this expression the asymptotical analysis of Section 5.2.3.4 can be applied. In the high SNR regime, we obtain,

$$\epsilon = \frac{Mt}{(Mt-1)}\frac{L}{P} + o\left(\frac{1}{P}\right). \tag{5.56}$$

Note that the distortion decay rate is 1, thus, confirming that linear analog approaches can not profit from either bandwidth expansion or diversity in terms of DDR. More interesting is the factor that precedes L/P in the first term on the right-hand side. This factor is a monotonically decreasing function of t, i.e., the benefit obtained from adding antennas at the receiver of the feedback link exceeds the disadvantage derived from having to feed back more channel coefficients.[10] This property of analog transmission has been pointed out in [81] in a different setting. In the low SNR regime, distortion behaves as

$$\epsilon_{\mathrm{OPTA}} = 1 - P/L + o(P).$$

This expression is the same as that given in Eq. 5.54 for the behavior of OPTA at low SNR values. Thus, also for this setting, linear analog transmission is optimum in the low SNR regime.

5.3.3 Numerical results

In this section we show numerical performance results of feedback links with multiple antennas at the receive end. The feedback channel is considered to be a time-dispersive channel with $M = 2$ delay taps and flat power delay profile. The coefficients of matrix $\boldsymbol{G}$ are the channel gains resulting from employing an OFDM transmission scheme over this

[10] Recall that derivation of this result has been done under the assumption $LMt \leq N$.

channel with $N = 16$ subcarriers. Fig. 5.8 shows curves for $t = 2$ antennas at the receive end of the feedback link. The slope of the OPTA curve at high SNR seems to be equal to 4, which is the bandwidth expansion factor in this setting. Elimination of time diversity results in an upper bound that seems to approach a slope of 4 at high SNR values. This indicates that the upper bounds on DDR derived in Section 5.3.1 might be tight. Comparing Figs. 5.6 and 5.8, we note that an additional antenna results in improvement of the theoretical upper bound without time diversity, the analog scheme and the digital schemes. Concerning the digital schemes, improvement can be noticed by observing that saturation is reached at lower SNR values. This improvement, predicted by our analysis in previous sections, is due to the increased diversity in the feedback link provided by the additional antenna. This is so despite the fact that, with one additional antenna, the number of coefficients that must be fed back in each use of the feedback link doubles from 2 to 4. Fig. 5.9 shows curves for the same setting with $t = 4$. Now, the bandwidth expansion factor decreases to 2. This is exactly the slope exhibited by the theoretical upper bounds at high SNR, which confirms that our upper bounds on DDR might be tight. The increase in diversity resulting from doubling the number of antennas yields some performance improvement for the analog scheme as predicted by Eq. 5.56. For the theoretical upper bounds and the digital schemes, the increase in diversity does not provide any benefit as, in this case, this is done at the cost of a reduced bandwidth expansion factor, which now becomes the limiting factor. As a consequence, a significant degradation of these curves can be observed.

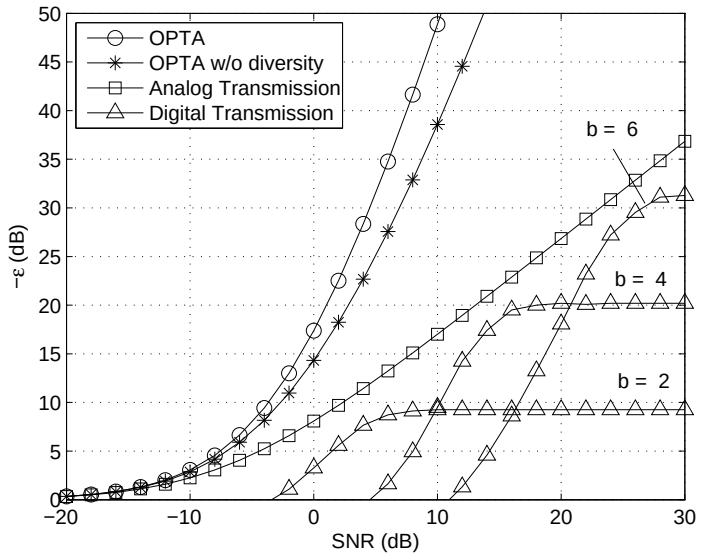

Figure 5.8: Performance over a fading feedback link with $L = 2$, $N = 16$, $M = 2$, $t = 2$.

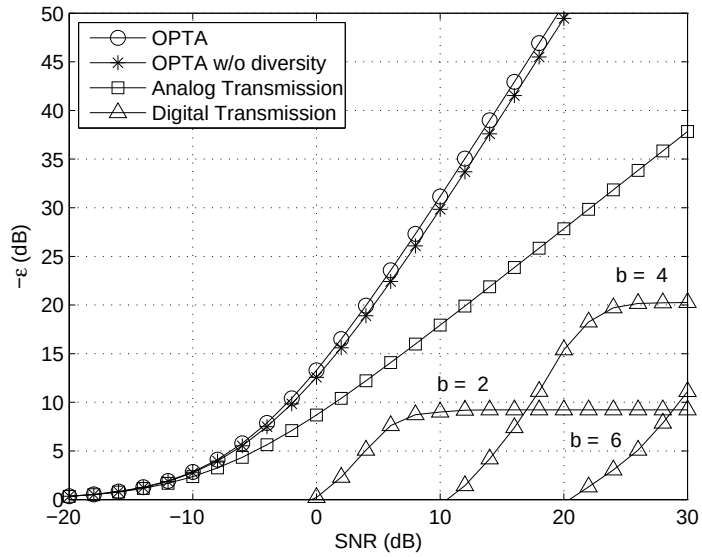

Figure 5.9: Performance over a fading feedback link with $L = 2$, $N = 16$, $M = 2$, $t = 4$.

5.4 Forward link performance under delay limited feedback

In this section, we consider performance of a broadcast channel in which receivers transmit information about the state of their respective channels back to the transmitter. A temporal block fading model is assumed according to which the channel state in the forward link changes abruptly from block to block and remains constant during the duration of one block. The receivers feed back CSI once per block. The transmitter processes the information received from all users and sets up the transmission parameters accordingly. No robust approach is considered. Instead, the transmitter uses the received CSI as if it were perfect. First, we shall briefly comment on the shortcomings of using ergodic rates as a performance measure when dirty paper coding is employed as a transmission scheme in the forward link. After identifying average throughput as a more appropriate measure, we present some numerical results.

5.4.1 Information theoretic measures

Most of the work on feedback schemes done so far uses ergodic capacity or ergodic capacity loss in the forward link as a performance measure, e.g., [78, 64, 20, 95]. A fundamental question that is largely neglected in the existing literature is how can the numerically computed ergodic capacity be effectively achieved. A closely related question is how does the transmitter know the maximum rate at which reliable transmission is possible. If a single-user forward link is considered, under the common error-free assumption on the feedback link, the receiver has the knowledge about both the real channel state and the imperfect information available at the transmitter. Thus, in such setting the receiver might compute

the maximum achievable rate and communicate it to the transmitter. If the time interval in which the channel state remains unchanged is long enough, the transmitter should be able to transmit at this rate reliably. In this case the ergodic capacity is achieved by employing different code books for different channel states and CSI. Obviously, if the assumption of error-free transmission on the feedback link is dropped, the receiver can not be sure about the CSI that the transmitter possesses any longer. Thus, neither the transmitter nor the receiver have enough information to compute the maximum rate achievable for a given channel state. However, even in this case, the ergodic capacity is achievable by transmitting with a unique code book over a large number of channel states with corresponding CSI. This can be shown by the following general reasoning from [24]. Let $p(y|x, u, T(u))$ be the probability transition function of the forward channel with output y and input x. Let u represent the state of the channel and $T(u)$ the transmit strategy, which is a function of the channel state. If u and $T(u)$ are viewed as additional channel outputs, the maximum achievable rate for an input with distribution $p(x)$ is given by $I(\mathsf{x}; \mathsf{y}, \mathsf{u}, T(\mathsf{u}))$, which can be achieved by a code book $\mathcal{C}_n$ of typical sequences x^n as $n \to \infty$. Using the chain rule for mutual information we can write

$$\begin{aligned} I(\mathsf{x}; \mathsf{y}, \mathsf{u}, T(\mathsf{u})) &= I(\mathsf{x}; \mathsf{y}|\mathsf{u}, T(\mathsf{u})) + I(\mathsf{x}; \mathsf{u}, T(\mathsf{u})) \\ &= \sum_{u} I(\mathsf{x}; \mathsf{y}|\mathsf{u} = u, T(\mathsf{u} = u))p(u), \end{aligned} \tag{5.57}$$

where it has been assumed that both the channel state and the transmit strategy are independent of the channel input and therefore $I(\mathsf{x}; \mathsf{u}, T(\mathsf{u})) = 0$. Note that the last expression is nothing else than the ergodic capacity for the given input statistics and transmit strategy. Thus, this capacity is achievable with a unique code book $\mathcal{C}_n$, $n \to \infty$. The transmitter can compute this rate provided that it has knowledge about the distribution of the channel states. Knowledge about the instantaneous channel states is not needed.

If a multiuser forward channel is considered, even under the assumption of a noiseless feedback link, the transmitter can not obtain information about the achievable rates from the receivers. This is due to the fact that the transmit strategy, and thus, the rates, will generally depend on the states of all single-user channels in the forward link and each receive terminal has only access to the own channel but not to the channels of other users. As in the single-user forward link, ergodic rates can be also achieved on the multiuser forward link provided that the input signals are independent of the channel states and the transmit strategy. This is true for linear approaches. To illustrate this, consider the two-user Gaussian broadcast channel given by

$$\boldsymbol{y}_1 = \boldsymbol{H}_1\boldsymbol{x} + \boldsymbol{n}_1,$$

$$\boldsymbol{y}_2 = \boldsymbol{H}_2\boldsymbol{x} + \boldsymbol{n}_2,$$

with $\boldsymbol{x} = \boldsymbol{B}_1\boldsymbol{s}_1 + \boldsymbol{B}_2\boldsymbol{s}_2$. Using the CSI fed back by the users, the transmitter computes beamforming matrices $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$, which represent the transmit strategy. For user 1, the effective channel can be written as

$$\boldsymbol{y}_1 = \boldsymbol{H}_1\boldsymbol{B}_1\boldsymbol{s}_1 + \boldsymbol{H}_1\boldsymbol{B}_2\boldsymbol{s}_2 + \boldsymbol{n}_1.$$

Assuming Gaussian inputs $\boldsymbol{s}_1$, $\boldsymbol{s}_2$ with the identity matrix as covariance matrix, the capacity of this channel can be written as $I(\boldsymbol{s}_1; \boldsymbol{y}_1, \boldsymbol{H}_1\boldsymbol{B}_1, \boldsymbol{H}_1\boldsymbol{B}_2\boldsymbol{B}_2^{\mathrm{H}}\boldsymbol{H}_1^{\mathrm{H}})$, where the product $\boldsymbol{H}_1\boldsymbol{B}_1$ and the covariance matrix $\boldsymbol{H}_1\boldsymbol{B}_2\boldsymbol{B}_2^{\mathrm{H}}\boldsymbol{H}_1^{\mathrm{H}}$ are viewed as channel outputs that the receiver has access to. There exist a code that allows reliable transmission at rates arbitrarily close to this mutual information as the length of the codewords approach to infinity. Furthermore, taking into account that the input $\boldsymbol{s}_1$ is independent of the channel matrices and the beamforming matrices, it can be shown as in Eq. 5.57 that this mutual information is equal to the ergodic capacity of this channel for the particular choice of input statistics and transmit strategy. The same reasoning applies to the rate achievable by user 2. In order to compute these rates, the transmitter only requires statistical knowledge of the channel states and transmit strategy, represented by the beamforming matrices.

Assume now that the users are encoded successively and a dirty paper coding scheme is employed based on the available CSI. Let $\hat{\boldsymbol{H}}_1$ and $\hat{\boldsymbol{H}}_2$ represent the CSI available at the transmitter. If user 2 is encoded first and $\boldsymbol{H}_1\boldsymbol{B}_1$ is invertible, the code book for user 1 is generated according to the random variable (cf. Section 2.2.2.2)

$$\boldsymbol{u}_1 = \boldsymbol{s}_1 + \boldsymbol{B}_1^{\mathrm{H}}\hat{\boldsymbol{H}}_1^{\mathrm{H}}\left(\boldsymbol{I}_{r_1} + \hat{\boldsymbol{H}}_1\boldsymbol{B}_1\boldsymbol{B}_1^{\mathrm{H}}\hat{\boldsymbol{H}}_1^{\mathrm{H}}\right)^{-1}\hat{\boldsymbol{H}}_1\boldsymbol{B}_2\boldsymbol{s}_2,$$

which obviously depends on the instantaneous channel state through the estimate $\hat{\boldsymbol{H}}_1$ and the beamforming matrices. That is, the code book is modified every time there is a change in the CSI. For particular channel states and particular channel estimates, the maximum rate achievable by user 1 can be computed as (cf. Sect 2.2.2.2)

$$R_1(\boldsymbol{H}_1, \boldsymbol{H}_2) = I(\boldsymbol{u}_1; \boldsymbol{y}_1) - I(\boldsymbol{u}_1; \boldsymbol{s}_2).$$

Averaging over all possible channel realizations an average rate can be computed. Different from the linear approaches, now, this rate is not achievable with a unique code book with codewords spanning a large number of different channel states. Furthermore, it is impossible for the transmitter to know the maximum rate achievable for instantaneous channel states and, therefore, this rate can neither be achieved by using different code books for different channel states and transmitting at the corresponding maximum instantaneous rate. Thus, if dirty paper coding is applied, the computed average rate does not have any theoretical meaning in the sense that it is actually unachievable. This seems to have been overlooked by recent publications on the topic, e.g., [43].

In this case, instead of ergodic rate, average throughput can be chosen as a figure of merit for the forward link. This figure is obtained by letting the transmitter fix a transmission rate for each single-user channel based on the CSI it receives from the users. If the actual rate supported by a particular single-user channel is larger than the transmission rate over that channel, transmission is declared successful and the instantaneous throughput is equal to the transmission rate. If the transmission rate is larger than the actual rate supported by the channel, the instantaneous throughput is considered to be zero. The average throughput is obtained by averaging instantaneous throughput over a large number of channel realizations. This figure captures the trade-off between the transmission rate guessed by the transmitter and the probability of this guess being too optimistic. The average throughput computed in this way can be effectively reached as long as the

time interval during which the channel state remains constant is long enough so that the probability of error becomes negligible at rates below the instantaneous capacity.

5.4.2 Numerical results

In this section, numerical results are presented for a broadcast channel with $K = 2$ users, $t = 2$ transmit antennas and single-antenna receivers. The forward and feedback channels of both users are assumed two be time-dispersive with two uncorrelated taps and a flat power delay profile, i.e., $L = M = 2$. In both forward and feedback links, an OFDM transmission scheme with $N = 16$ subcarriers is used. After perfect estimation of the state of the single-user channels, the receivers feed the channel coefficients back to the transmitter in a time-division multiple access fashion. Each user utilizes one OFDM symbol, i.e., $N = 16$ subcarriers, for the transmission of the 4 channel coefficients representing the estimated channel state. Upon reception of the channel coefficients of both users, the transmitter executes the sum-rate maximizing SESAM algorithm (cf. Section 4.1.4) and transmission is carried out by using the resulting beamforming vectors and power loading. For both foward and feedback links SNR $= P/N$ is assumed to be equal. In Fig. 5.10 average throughput curves are shown that are obtained if the transmitter fully relies on the received CSI in order to determine the transmission rate on each subchannel. If the CSI information is perfect, the average sum-throughput coincides with the average sum-rate measure. In such case, the transmitter perfectly knows the maximum rate supported by the different subchannels and no transmission failures occur. If the CSI is obtained from a noisy feedback link and the transmission rate is chosen to be that computed based on this information, performance degrades dramatically. The reason for this degradation is that the estimated maximum achievable rate on a particular subchannel frequently exceeds the actual transmission rate supported by that subchannel resulting in failed transmissions. As a simple countermeasure, on each subchannel the transmission rate can be set a certain margin below the rate computed based on the imperfect CSI. Fig. 5.11 shows performance curves obtained when the transmitter sets the transmission rate on each subchannel 0.5 bits per channel use below the estimated maximum achievable rate. The performance improvement is notorious. The gap between the average throughput obtained with perfect CSI and that resulting from CSI fed back using a simple linear analog approach is slightly above 1 bit. In all the range of simulated SNR values and for all practical purposes, digital schemes do at best perform as well as the analog scheme. The gap between perfect CSI and analog feedback CSI being so narrow, it is difficult to imagine how to turn parameters so that a clear superiority of delay-limited digital schemes becomes visible. Keeping the forward link fixed, if the bandwidth expansion factor and the diversity degree in the feedback link are increased, performance of the digital schemes is expected to improve. However, if not in terms of DDR, the analog scheme will also profit from such a change, which would make the performance gap with respect to the perfect CSI curve narrower. On the other hand, if both diversity and bandwidth expansion factor decrease in the feedback link, performance of digital schemes will degrade more significantly than performance of the analog scheme, making digital approaches less competitive.

Linear analog transmission is a very simple scheme that, as we have seen in previous sections, has important performance limitations. However, in the light of these numerical

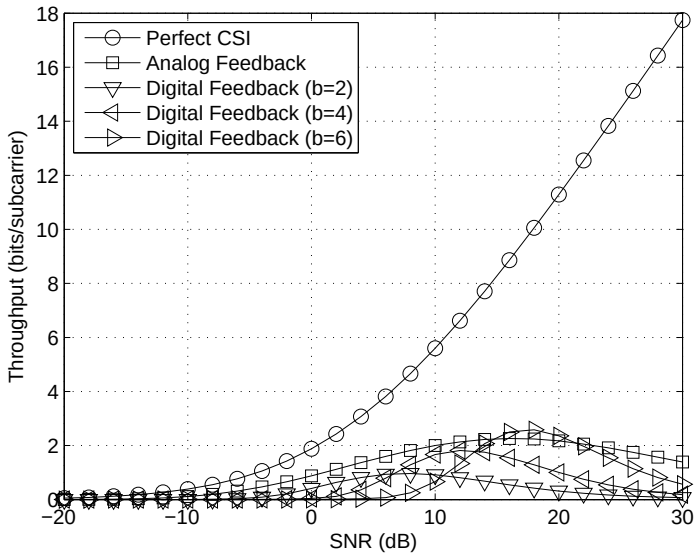

Figure 5.10: Achievable sum throughput in a broadcast forward link with $N = 16$ subcarriers, $K = 2$ users, $t = 2$ transmit antennas and single-antenna receivers. Transmitter fixes the transmission rate as if the CSI were perfect.

results, we may conjecture that this scheme is good enough for purposes of feedback of CSI. As we have seen, in general, digital schemes have the theoretical potential to perform better. However, even if digital schemes are found that at affordable complexity yield a better performance in terms of MSE at SNR values of interest, when translated into forward link performance, the resulting gains might be insignificant.

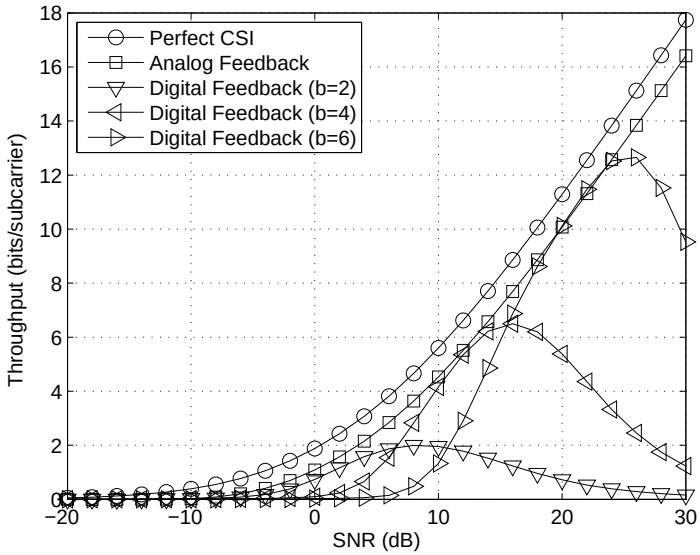

Figure 5.11: Achievable sum throughput in a broadcast forward link with $N = 16$ subcarriers, $K = 2$ users, $t = 2$ transmit antennas and single-antenna receivers. Transmitter allows for a rate margin of 0.5 bits per subchannel.

A Appendix

A.1 Duality transformations and the matrix inversion lemma

A.1.1 Duality transformations

Recall the model for the broadcast channel given by Eq. 2.6 and its dual multiple access channel given in Eq. 2.25. Without loss of generality, assume that, in the MAC, users are decoded in the order indicated by their indexes, and, for this channel, let $\boldsymbol{Q}_{1,\ldots,K}$ be a set of transmit covariance matrices. Define

$$\boldsymbol{B}_j = \boldsymbol{I}_t + \sum_{k=j+1}^{K} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{Q}_k \boldsymbol{H}_k,\ j = 1, \ldots, K,$$

and

$$\boldsymbol{A}_j = \boldsymbol{I}_{r_j} + \boldsymbol{H}_j \left(\sum_{k=1}^{j-1} \boldsymbol{\Sigma}_k \right) \boldsymbol{H}_j^{\mathrm{H}},\ j = 1, \ldots, K,$$

and consider the singular value decomposition (SVD)

$$\boldsymbol{B}_k^{-1/2} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{A}_k^{-1/2} = \boldsymbol{F}_k \boldsymbol{\Lambda}_k \boldsymbol{G}_k^{\mathrm{H}},\ k = 1, \ldots, K.$$

The transformations

$$\boldsymbol{\Sigma}_k = \boldsymbol{B}_k^{-1/2} \boldsymbol{F}_k \boldsymbol{G}_k^{\mathrm{H}} \boldsymbol{A}_k^{1/2} \boldsymbol{Q}_k \boldsymbol{A}_k^{1/2} \boldsymbol{G}_k \boldsymbol{F}_k^{\mathrm{H}} \boldsymbol{B}_k^{-1/2},\ k = 1, \ldots, K$$

provide a set $\boldsymbol{\Sigma}_{1,\ldots,K}$ of transmit covariance matrices for the BC that reach the same rate vector as the matrices $\boldsymbol{Q}_{1,\ldots,K}$ in the MAC, provided that the encoding order in the BC corresponds to the reversed decoding order in the MAC, i.e., user K is encoded first and user 1 last.

A.1.2 Matrix inversion lemma

Let $\boldsymbol{A}$ and $\boldsymbol{C}$ be invertible square matrices of dimensions $N \times N$ and $L \times L$, respectively. Further, let $\boldsymbol{V}$ be an $L \times N$ matrix and $\boldsymbol{U}$ an $N \times L$ matrix. The following equality holds,

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{C}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{C}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}.$$

A.2 Asymptotic equipartition property and typical sequences

In this appendix an overview on the subject of typical sequences and their properties is given. This introduction is intended to help the reader to better understand the arguments used in the achievability proofs of some of the results presented in Chapter 2. For clarity of exposition we assume discrete alphabets. However, all definitions and theorems can be trivially extended to the case of continuous alphabets by replacing entropies with differential entropies and cardinalities of typical sets with volumes of these sets. For a more detailed discussion on the topic the reader is referred to [40].

Theorem A.2.1 (Asymptotic Equipartition Property)**.** *If $\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_n$ are independent and identically distributed random variables drawn according to a distribution $p(x)$, then for $n \to \infty$*

$$-\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) \to H(\mathsf{x}),$$

where $H(\mathsf{x}) = -\mathrm{E}\{\log p(\mathsf{x})\}$ is the entropy of x.

Proof. Functions of independent random variables are also independent random variables. In particular, since $\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_n$ are i.i.d., $\log p(\mathsf{x}_1), \log p(\mathsf{x}_2), \ldots, \log p(\mathsf{x}_n)$ are also i.i.d.. As a result, the weak law of large numbers [89] can be applied in order to show

$$-\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) = -\frac{1}{n}\sum_{i=1}^{n}\log p(x_i) \to -\mathrm{E}\{\log p(\mathsf{x})\} = H(\mathsf{x}),$$

for $n \to \infty$. □

The typical set $A_\epsilon^{(n)}$ with respect to $p(x)$ is defined as the set of sequences $x^n \in \mathcal{X}^n$ with the property

$$2^{-n(H(\mathsf{x})+\epsilon)} < p(x_1, x_2, \ldots, x_n) < 2^{-n(H(\mathsf{x})-\epsilon)}.$$

The sequences that belong to this set are said to be typical. The typical set has the following properties.

Theorem A.2.2 (Properties of the typical set)**.**

1. $Pr\left\{A_\epsilon^{(n)}\right\} > 1-\epsilon$ *for sufficiently large* n.
2. $\left|A_\epsilon^{(n)}\right| \leq 2^{n(H(\mathsf{x})+\epsilon)}$.
3. $\left|A_\epsilon^{(n)}\right| \geq (1-\epsilon)2^{n(H(\mathsf{x})-\epsilon)}$ *for sufficiently large* n.

Proof. Let x^n be a sequence generated according to $p(x)$, then

$$Pr\left\{A_\epsilon^{(n)}\right\} = Pr\left\{x^n \in A_\epsilon^{(n)}\right\} = Pr\left\{\left|-\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) - H(\mathsf{x})\right| < \epsilon\right\}$$

which due to the asymptotic equipartition property tends to 1 as $n \to \infty$. This proves the first property. In order to prove the second property consider the inequalities,

$$1 \geq \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) \geq \sum_{x^n \in A_\epsilon^{(n)}} 2^{-n(H(\mathsf{x})+\epsilon)} = \left|A_\epsilon^{(n)}\right| 2^{-n(H(\mathsf{x})+\epsilon)},$$

from which $\left|A_\epsilon^{(n)}\right| \leq 2^{n(H(\mathsf{x})+\epsilon)}$ follows. Similarly, invoking property 1, for sufficiently large n we can write

$$1 - \epsilon < \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) \leq \sum_{x^n \in A_\epsilon^{(n)}} 2^{-n(H(\mathsf{x})-\epsilon)} = \left|A_\epsilon^{(n)}\right| 2^{-n(H(\mathsf{x})-\epsilon)},$$

which proves the third property. □

The first property tells us that in the limit $n \to \infty$ all sequences generated according to a distribution $p(x)$ are typical with respect to that distribution. The second and third property tell us that as $n \to \infty$ the number of typical sequences approaches to $2^{nH(\mathsf{x})}$.

The asymptotic equipartition property and the concept of typicality can be extended to distributions of multiple variables. Consider two random variables x and y that are distributed according to $p(x, y)$. The set $A_\epsilon^{(n)}$ of jointly typical sequences (x^n, y^n) with respect to this distribution is defined as

$$\begin{aligned} A_\epsilon^{(n)} = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \; & \left|-\frac{1}{n}\log p(x^n) - H(\mathsf{x})\right| < \epsilon, \\ & \left|-\frac{1}{n}\log p(y^n) - H(\mathsf{y})\right| < \epsilon, \\ & \left|-\frac{1}{n}\log p(x^n, y^n) - H(\mathsf{x}, \mathsf{y})\right| < \epsilon\right\}. \end{aligned}$$

That is, a pair of sequences (x^n, y^n) is jointly typical if each sequence individually is typical and they also behave typically with respect to each other. Jointly typical sequences have the following properties.

Theorem A.2.3 (Properties of jointly typical sequences)**.** *Let (x^n, y^n) be sequences of length n drawn according to $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$. Then*

1. $Pr\left\{(x^n, y^n) \in A_\epsilon^{(n)}\right\} > 1 - \epsilon$ *for sufficiently large n.*
2. $\left|A_\epsilon^{(n)}\right| \leq 2^{n(H(\mathsf{x},\mathsf{y})+\epsilon)}$.
3. $\left|A_\epsilon^{(n)}\right| \geq (1-\epsilon)2^{n(H(\mathsf{x},\mathsf{y})-\epsilon)}$ *for sufficiently large n.*
4. *If $(\tilde{x}^n, \tilde{y}^n)$ are generated according to $p(x)p(y)$, i.e., the sequences are independently generated according to the marginals of $p(x, y)$, then*

$$Pr\left\{(\tilde{x}^n, \tilde{y}^n) \in A_\epsilon^{(n)}\right\} \leq 2^{-n(I(\mathsf{x},\mathsf{y})-3\epsilon)}.$$

Also, for sufficiently large n,

$$Pr\left\{(\tilde{x}^n, \tilde{y}^n) \in A_\epsilon^{(n)}\right\} \geq (1-\epsilon)2^{-n(I(\mathsf{x},\mathsf{y})+3\epsilon)}.$$

Proof. The proof of the first three properties can be conducted along the lines of the proof of Theorem A.2.2. In order to prove the first part in property 4, we consider the following inequalities

$$\begin{aligned} Pr\left\{(\tilde{x}^n, \tilde{y}^n) \in A_\epsilon^{(n)}\right\} &= \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} p(x^n)p(y^n) \leq \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} 2^{-n(H(x)-\epsilon)}2^{-n(H(y)-\epsilon)} \\ &= \left|A_\epsilon^{(n)}\right| 2^{-n(H(x)-\epsilon)}2^{-n(H(y)-\epsilon)} \leq 2^{n(H(x,y)+\epsilon)}2^{-n(H(x)-\epsilon)}2^{-n(H(y)-\epsilon)} \\ &= 2^{-n(I(x,y)-3\epsilon)}. \end{aligned}$$

Similarly, we can prove the second part by writing

$$\begin{aligned} Pr\left\{(\tilde{x}^n, \tilde{y}^n) \in A_\epsilon^{(n)}\right\} &= \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} p(x^n)p(y^n) \geq \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} 2^{-n(H(x)+\epsilon)}2^{-n(H(y)+\epsilon)} \\ &= \left|A_\epsilon^{(n)}\right| 2^{-n(H(x)-\epsilon)}2^{-n(H(y)-\epsilon)} \geq (1-\epsilon)2^{n(H(x,y)+\epsilon)}2^{-n(H(x)-\epsilon)}2^{-n(H(y)-\epsilon)} \\ &= (1-\epsilon)2^{-n(I(x,y)+3\epsilon)}. \end{aligned}$$

Note that property 3 has been invoked in the last inequality and this property only holds for sufficiently large n. □

All the achievability proofs discussed in Chapter 2 assume receivers that perform detection based on joint typicality. Given a received sequence of signals, the receiver looks for a sequence in the code book that is jointly typical with that received sequence. Detection is successful if the solution exists, i.e., a codeword can be found that is jointly typical with the received sequence, and this solution is unique, i.e., there is only one such sequence. Property 1 in Theorem A.2.3 guarantees that if the typical sequence x^n is transmitted over a channel with transition probability $p(y|x)$ the received sequence y^n is jointly typical with x^n with probability tending to one as $n \to \infty$. That is, the first property guarantees that, for large n, at least the transmitted codeword will be jointly typical with the received signal. Uniqueness holds if the probability that x^n and y^n be jointly typical is zero when x^n is not the codeword that gave rise to the received sequence y^n, i.e., when x^n and y^n where independently generated. Assume a code book $\mathcal{C}_n$ consisting of 2^{nR} typical sequences x^n and that transmission of the sequence x_1^n leads to the received sequence y^n. Using the first part of property 4 in Theorem A.2.3 we can bound the probability of any other $x_{i\neq 1}^n$ being jointly typical with y^n as follows,

$$Pr\left\{\exists\ x_{i\neq 1}^n : (x_i^n, y^n) \in A_\epsilon^{(n)}\right\} \leq \sum_{i=2}^{2^{nR}} Pr\left\{(x_i^n, y^n) \in A_\epsilon^{(n)}\right\} \leq 2^{nR}2^{-n(I(x,y)-3\epsilon)}.$$

Thus, if transmission rate is chosen such that $R < I(x, y)$ the probability that uniqueness does not hold tends to zero as $n \to \infty$.

In some of the proofs, in order to assure typicality of the transmitted signals it is required that pairs of jointly typical sequences exist within a set that was created by generating sequences independently. For instance, this is needed in the achievability proof of the Marton's region for both the simultaneous and the successive encoding schemes (cf. Sections 2.1.3.1, 2.2.2.2). Assume that 2^{nR_1} sequences u^n and 2^{nR_2} sequences v^n are independently

generated according to distributions $p(u)$ and $p(v)$, respectively. Given a joint probability density function $p(u,v)$, the probability that a jointly typical pair can be found among the set of $2^{n(R_1+R_2)}$ pairs of sequences (u^n, v^n) can be, for sufficiently large n, lower bounded as follows

$$\begin{aligned} Pr\left\{\exists\ (i,j) \in \{1,\ldots,2^{nR_1}\} \times \{1,\ldots,2^{nR_2}\} : (u_i^n, v_j^n) \in A_\epsilon^{(n)}\right\} \\ &= 1 - Pr\left\{\forall\ (i,j) \in \{1,\ldots,2^{nR_1}\} \times \{1,\ldots,2^{nR_2}\} : (u_i^n, v_j^n) \notin A_\epsilon^{(n)}\right\} \\ &= 1 - \prod_{i,j} Pr\left\{(u_i^n, v_j^n) \notin A_\epsilon^{(n)}\right\} \\ &\geq 1 - \left(1 - (1-\epsilon)2^{-n(I(u,v)+3\epsilon)}\right)^{2^{n(R_1+R_2)}} \\ &\geq 1 - \exp -(1-\epsilon)2^{n(R_1+R_2)}2^{-n(I(u,v)+3\epsilon)} \end{aligned}$$

The first inequality is due to the second part of property 4 in Theorem A.2.3. In order to obtain the second inequality the relation $(1-x)^n \leq \exp -nx$, for $x < 1$, $n \geq 0$, has been used. Thus, we note that if $R_1 + R_2 > I(\mathsf{u}, \mathsf{v})$ the probability that a jointly typical pair can be found tends to 1 as $n \to \infty$.

A.3 Langrangian duality and subgradients

In this appendix a brief overview on basic optimization theoretic results, mostly used in Chapter 3, is given. To a large extend, the exposition is based on [13]. In some parts of the text, additional references have also been included to sources that exhibit complementary points of view in the treatment of these contents.

Consider the following optimization problem

$$\begin{aligned} &\max_{\boldsymbol{Q}}\ f(\boldsymbol{Q}), \\ &\text{subject to} \quad \boldsymbol{Q} \geq \mathbf{0},\ \ h(\boldsymbol{Q}) \geq \mathbf{0}, \end{aligned} \tag{A.1}$$

where $\boldsymbol{Q} \in \mathbb{H}^{n\times n}$, $f : \mathbb{H}^{n\times n} \to \mathbb{R}$ and $h : \mathbb{H}^{n\times n} \to \mathbb{R}$. In the following, unless otherwise stated, the discussion will be based on this example. While keeping the treatment at a general level, this will allow us to capitalize on particular details of the theory relevant to the specific problems analyzed in Chapter 3. One such detail is the fact that optimization is performed over the set of Hermitian matrices with a positive semidefinite constraint. Henceforth, problem A.1 will be referred to as primal problem.

A.3.1 Lagrangian duality

For problem A.1 the Lagrangian function is given by

$$L(\boldsymbol{Q}, \lambda, \boldsymbol{\Phi}) = f(\boldsymbol{Q}) + \lambda h(\boldsymbol{Q}) + \mathrm{Tr}\left\{\boldsymbol{Q}\boldsymbol{\Phi}\right\},$$

where $\lambda \in \mathbb{R}$ and $\boldsymbol{\Phi} \in \mathbb{H}^{n\times n}$ are called the Lagrangian multipliers associated with constraints $h(\boldsymbol{Q}) \geq \mathbf{0}$ and $\boldsymbol{Q} \geq \mathbf{0}$, respectively. A major property of this function is that

$$f(\boldsymbol{Q}) \leq L(\boldsymbol{Q}, \lambda, \boldsymbol{\Phi}), \quad \forall\ \lambda \geq 0,\ \ \forall\ \boldsymbol{\Phi} \geq \mathbf{0}, \tag{A.2}$$

for every feasible $\boldsymbol{Q}$, i.e., for every $\boldsymbol{Q}$ that satisfies the constraints in problem A.1. That is, if the multipliers are nonnegative, the Lagrangian function upper bounds the objective function for all points of the feasibility region. Based on the Lagrangian function, the Lagrangian dual function is defined as

$$g(\lambda, \boldsymbol{\Phi}) = \sup \left\{ L(\boldsymbol{Q}, \lambda, \boldsymbol{\Phi}) \mid \boldsymbol{Q} \in \mathbb{H}^{n \times n} \right\}.$$

As a result of Eq. A.2,

$$f(\boldsymbol{Q}) \leq g(\lambda, \boldsymbol{\Phi}), \quad \forall\ \lambda \geq 0, \quad \forall\ \boldsymbol{\Phi} \geq \mathbf{0}, \tag{A.3}$$

for every feasible $\boldsymbol{Q}$. Furthermore, for $\mu_1, \mu_2 \geq 0$, $\mu_1 + \mu_2 = 1$,

$$\begin{aligned}
\mu_1 g(\lambda_1, \boldsymbol{\Phi}_1) \ &+ \ \mu_2 g(\lambda_2, \boldsymbol{\Phi}_2) = \\
&= \ \sup \left\{ \mu_1 L(\boldsymbol{Q}, \lambda_1, \boldsymbol{\Phi}_1) \mid \boldsymbol{Q} \in \mathbb{H}^{n \times n} \right\} + \sup \left\{ \mu_2 L(\boldsymbol{Q}, \lambda_2, \boldsymbol{\Phi}_2) \mid \boldsymbol{Q} \in \mathbb{H}^{n \times n} \right\} \\
&\geq \ \sup \left\{ \mu_1 L(\boldsymbol{Q}, \lambda_1, \boldsymbol{\Phi}_1) + \mu_2 L(\boldsymbol{Q}, \lambda_2, \boldsymbol{\Phi}_2) \mid \boldsymbol{Q} \in \mathbb{H}^{n \times n} \right\} \\
&= \ \sup \left\{ L(\boldsymbol{Q}, \mu_1\lambda_1 + \mu_2\lambda_2, \mu_1\boldsymbol{\Phi}_1 + \mu_2\boldsymbol{\Phi}_2) \mid \boldsymbol{Q} \in \mathbb{H}^{n \times n} \right\} \\
&= \ g(\mu_1\lambda_1 + \mu_2\lambda_2, \mu_1\boldsymbol{\Phi}_1 + \mu_2\boldsymbol{\Phi}_2),
\end{aligned}$$

i.e., $g(\lambda, \boldsymbol{\Phi})$ is convex. So far the Lagrangian function and the Lagrangian dual function have been defined with respect to both constraints in the primal problem. Sometimes, it is useful to define these function with respect to a subset of constraints (cf. Section 3.1.1.2). For instance, we could have defined

$$L(\boldsymbol{Q}, \lambda) = f(\boldsymbol{Q}) + \lambda h(\boldsymbol{Q}), \tag{A.4}$$

$$g(\lambda) = \sup \left\{ L(\boldsymbol{Q}, \lambda) \mid \boldsymbol{Q} \in \mathbb{H}^{n \times n},\ \boldsymbol{Q} \geq \mathbf{0} \right\}. \tag{A.5}$$

It can be easily shown that $g(\lambda)$ is convex and both functions upper bound $f(\boldsymbol{Q})$ in the feasible domain as long as $\lambda \geq 0$.

If all constraints are considered in the definition of the Lagrangian and the Lagrangian dual functions, the dual problem corresponding to the primal problem A.1 can be stated as

$$\begin{aligned}
&\min_{\boldsymbol{\Phi}, \lambda} \ g(\lambda, \boldsymbol{\Phi}), \\
&\text{subject to} \quad \lambda \geq \mathbf{0}, \quad \boldsymbol{\Phi} \geq \mathbf{0}.
\end{aligned} \tag{A.6}$$

If the definitions given by Eqs. A.4 and A.5 are considered, the dual problem simplifies to

$$\begin{aligned}
&\min_{\lambda} \ g(\lambda), \\
&\text{subject to} \quad \lambda \geq 0.
\end{aligned} \tag{A.7}$$

Note that due to convexity of the objective functions and constraints in Eqs. A.6 and A.7, both dual problems are convex. Let $\bar{f}$ and $\bar{g}$ denote the solutions of the primal and any of the dual problems, respectively. Due to Eq. A.3, which also holds for $g(\lambda)$, $\bar{f} \leq \bar{g}$. This result is known as weak duality and states that the solutions of the dual problems upper

bound the solution of the primal problem. Strong duality holds if $\bar{f} = \bar{g}$. It can be easily shown that the solution to problem A.6 upper bounds the solution to problem A.7. Thus, if strong duality holds for the dual problem defined with respect to all constraints, it also holds for the dual problem defined with respect to a reduced number of constraints. For a primal optimization problem with inequality constraints, as the one considered here, strong duality holds if the problem is convex and there exists a feasible point that satisfies all the constraints with strict inequality. This sufficient condition is related to Slater's constraint qualification [13, 4].

A.3.2 Optimality conditions

Assume that strong duality holds for problem A.1 and let $\bar{\lambda}$, $\bar{\boldsymbol{\Phi}}$ be the minimizers of problem A.6 and $\bar{\boldsymbol{Q}}$ the maximizer of problem A.1. The following holds

$$f(\bar{\boldsymbol{Q}}) = g(\bar{\lambda}, \bar{\boldsymbol{\Phi}}) \geq L(\bar{\boldsymbol{Q}}, \bar{\lambda}, \bar{\boldsymbol{\Phi}}) \geq f(\bar{\boldsymbol{Q}}). \tag{A.8}$$

Thus, it can be observed that the inequalities in this expression must be satisfied with equality. In order to achieve equality in the second inequality we need

$$\lambda h(\boldsymbol{Q}) = 0, \quad \mathrm{Tr}\left\{\boldsymbol{Q}\boldsymbol{\Phi}\right\} = 0. \tag{A.9}$$

These conditions are known as complementary slackness. Achieving equality in the first inequality of Eq. A.8 requires that $\bar{\boldsymbol{Q}}$ be a maximizer of $L(\boldsymbol{Q}, \bar{\lambda}, \bar{\boldsymbol{\Phi}})$. Assuming differentiability of both $f(\boldsymbol{Q})$ and $h(\boldsymbol{Q})$ at $\bar{\boldsymbol{Q}}$, this implies

$$\nabla_{\boldsymbol{Q}} L(\bar{\boldsymbol{Q}}, \bar{\lambda}, \bar{\boldsymbol{\Phi}}) = \boldsymbol{0}. \tag{A.10}$$

Eqs. A.10 and A.9 together with the constraints in the primal and dual problems form a set of necessary optimality conditions for a wide range of problems with differentiable objective and constraint functions. These conditions are frequently referred to as Karush-Kuhn-Tucker (KKT) conditions. If the primal problem is convex, these conditions are sufficient. Note that the above derivation of Eqs. A.10 and A.9 assumes strong duality. This assumption is not necessary for the derivation of the KKT conditions. For alternative derivations that do not assume strong duality see for instance [4, 7, 94].

A.3.3 Subgradients

Lagrangian duality offers the possibility to find the solution to a given problem or a bound of that solution by stating and solving its corresponding dual. As we have seen, dual problems are always convex. However, due to the way it is defined, the objective function of a dual problem might not be differentiable. In that case, algorithmic solutions are based on the notion of subgradient rather than that of gradient. Consider a possibly non-differentiable function $g : \mathbb{R}^n \to \mathbb{R}$. A subgradient for this function at $\boldsymbol{\lambda}$ is a vector $\boldsymbol{s}$ that satisfies

$$g(\boldsymbol{\lambda} + \boldsymbol{\Delta\lambda}) \geq g(\boldsymbol{\lambda}) + \boldsymbol{s}^{\mathrm{T}}\boldsymbol{\Delta\lambda}, \quad \forall \boldsymbol{\Delta\lambda}. \tag{A.11}$$

If the function $g(\boldsymbol{\lambda})$ is differentiable at a particular point, the only vector $\boldsymbol{s}$ that satisfies Eq. A.11 is the gradient of this function at that point. In other words, if differentiability holds,

the concepts of gradient and subgradient are equivalent. The definition of subgradient given above can be generalized in order to include functions with Hermitian matrices as arguments (cf. Eq. A.6). Here, we omit this definition as it shall not find application in this work. Subgradients of Lagrangian dual functions can be straightforwardly found. In general, these are given by the constraint functions of the primal problem with respect to which the dual function is defined evaluated at the Lagrangian maximizing primal variables. As an example, consider problem A.7. A subgradient for the objective function at λ is simply $h(\bar{\boldsymbol{Q}})$, where $\bar{\boldsymbol{Q}}$ is a maximizer of $L(\boldsymbol{Q}, \lambda)$ subject to $\boldsymbol{Q} \geq \boldsymbol{0}$. An introduction to subgradient-based optimization methods can be found in [14, 7].

A.4 Duality of streamwise multiuser strategies

This appendix describes the duality relationship between approaches in the broadcast channel and the multiple access channel that are based on streamwise transmission. By streamwise, we refer to strategies that decompose the MIMO broadcast channel into a set of scalar subchannels over which streams of information can be transmitted that are independently encoded and decoded. This is in contrast to non-streamwise approaches, where the information intended for a particular user is conveyed by a vector of signals whose components are encoded and decoded jointly. In the first section of this appendix, it is shown that streamwise transmission does not incur loss of optimality in the MAC, i.e., streamwise approaches achieve all points of the capacity region. In the second section we will show a streamwise duality between BC and MAC. On the one hand, this duality underlies the derivation of the streamwise SINR-based successive approach presented in Section 4.2. On the other hand, it constitutes the proof for the optimality of streamwise transmission in the broadcast channel.

A.4.1 Optimality of streamwise strategies

Recall the Gaussian multiple access channel definition in Eq. 2.25, i.e.,

$$\boldsymbol{r} = \sum_{k=1}^{K} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{w}_k + \boldsymbol{z}.$$

Further, let $\boldsymbol{Q}_k$ be the covariance matrix corresponding to user k and $\boldsymbol{Q}_k = \boldsymbol{F}_k \boldsymbol{F}_k^{\mathrm{H}}$ be a factorization of this matrix. Denote by $\boldsymbol{f}_{k,j}$ the jth column of $\boldsymbol{F}_k$. According to these definitions, the signal transmitted by user k can be written as

$$\boldsymbol{w}_k = \sum_{j=1}^{m_k} \boldsymbol{f}_{k,j} s_{k,j},$$

where m_k can be viewed as the number of streams transmitted by user k and $s_{k,j}$ is the signal transmitted on the jth stream of user k. Signals $s_{k,j}$, $k = 1, \ldots, K$, $j = 1, \ldots, m_k$, are considered to be realizations of mutually independent zero-mean and unit-variance circularly symmetric complex Gaussian random variables. Thus, we see that, for any given

statistics, the vector of transmit signals corresponding to any particular user can be decomposed into a superposition of independent scalar transmit signals or streams that are transmitted by employing corresponding beamforming vectors. In the following, we shall see than independent detection of these streams based on a minimum mean squared error decision-feedback equalizer (MMSE-DFE) preserves capacity. To this end, assume that users are successively decoded in the order indicated by their indexes, i.e., $\bar{\pi}(k) = k$, and consider detection of the signals corresponding to user k. Provided that the streams of users $1, \ldots, k-1$ are detected without errors, the contributions of these streams to the received signal can be removed before detecting the streams of user k. The resulting received signal is given by

$$\boldsymbol{r}_k^{(1)} = \boldsymbol{H}_k^{\mathrm{H}} \sum_{j=1}^{m_k} \boldsymbol{f}_{k,j} s_{k,j} + \hat{\boldsymbol{z}},$$

where $\hat{\boldsymbol{z}}$ is a term of interference plus noise including the received signals corresponding to users $k+1, \ldots, K$. Let $\boldsymbol{R}_{\hat{\boldsymbol{z}}}$ be the covariance matrix of this term. The maximum rate achievable by user k can be written as

$$I(\mathsf{s}_{k,1}, \ldots, \mathsf{s}_{k,m_k}; \boldsymbol{r}_k^{(1)}) = \log_2 \left(\frac{|\boldsymbol{R}_{\hat{\boldsymbol{z}}} + \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{Q}_k \boldsymbol{H}_k|}{|\boldsymbol{R}_{\hat{\boldsymbol{z}}}|} \right).$$

Applying the chain rule for mutual information, we can write

$$I(\mathsf{s}_{k,1}, \ldots, \mathsf{s}_{k,m_k}; \boldsymbol{r}_k^{(1)}) = I(\mathsf{s}_{k,1}; \boldsymbol{r}_k^{(1)}) + I(\mathsf{s}_{k,2}, \ldots, \mathsf{s}_{k,m_k}; \boldsymbol{r}_k^{(1)} | \mathsf{s}_{k,1}).$$

The first term can be computed as

$$I(\mathsf{s}_{k,1}; \boldsymbol{r}_k^{(1)}) = h(\mathsf{s}_{k,1}) - h(\mathsf{s}_{k,1} | \boldsymbol{r}_k^{(1)}) = \log_2 \left(\frac{1}{\sigma^2_{\mathsf{s}_{k,1} | \boldsymbol{r}_k^{(1)}}} \right),$$

where $\sigma^2_{\mathsf{s}_{k,1} | \boldsymbol{r}_k^{(1)}}$ is the variance of the distribution $p(\mathsf{s}_{k,1} | \boldsymbol{r}_k^{(1)})$. This variance is independent of $\boldsymbol{r}_k^{(1)}$ and represents the minimum mean squared error that can be achieved if the estimation of $\mathsf{s}_{k,1}$ is based on $\boldsymbol{r}_k^{(1)}$ (cf. [144]). Let $\hat{\mathsf{s}}_{k,1}$ be the MMSE estimate of $\mathsf{s}_{k,1}$. Noting that $\mathsf{s}_{k,1} = \hat{\mathsf{s}}_{k,1} + \mathsf{e}_{k,1}$ and that the error $\mathsf{e}_{k,1}$ is independent of $\hat{\mathsf{s}}_{k,1}$ we can write

$$I(\mathsf{s}_{k,1}; \hat{\mathsf{s}}_{k,1}) = h(\mathsf{s}_{k,1}) - h(\mathsf{s}_{k,1} | \hat{\mathsf{s}}_{k,1}) = h(\mathsf{s}_{k,1}) - h(\mathsf{e}_{k,1}) = \log_2 \left(\frac{1}{\sigma^2_{\mathsf{s}_{k,1} | \boldsymbol{r}_k^{(1)}}} \right),$$

i.e., $I(\mathsf{s}_{k,1}; \boldsymbol{r}_k^{(1)}) = I(\mathsf{s}_{k,1}; \hat{\mathsf{s}}_{k,1})$. For the second term we can write

$$\begin{aligned} I(\mathsf{s}_{k,2}, \ldots, \mathsf{s}_{k,m_k}; \boldsymbol{r}_k^{(1)} | \mathsf{s}_{k,1}) &= h(\boldsymbol{r}_k^{(1)} | \mathsf{s}_{k,1}) - h(\boldsymbol{r}_k^{(1)} | \mathsf{s}_{k,1}, \mathsf{s}_{k,2}, \ldots, \mathsf{s}_{k,m_k}) \\ &= h(\boldsymbol{r}_k^{(2)}) - h(\boldsymbol{r}_k^{(2)} | \mathsf{s}_{k,2}, \ldots, \mathsf{s}_{k,m_k}) \\ &= I(\mathsf{s}_{k,2}, \ldots, \mathsf{s}_{k,m_k}; \boldsymbol{r}_k^{(2)}), \end{aligned}$$

where $\boldsymbol{r}_k^{(2)} = \boldsymbol{H}_k^{\mathrm{H}} \sum_{j=2}^{m_k} \boldsymbol{f}_{k,j} s_{k,j} + \hat{\boldsymbol{z}}$. The first term can be viewed as the maximum rate that can be transmitted over the stream $s_{k,1}$ if an MMSE estimate of this stream is taken

at the receiver. The second term represents the information rate of all other streams when the contribution of the first stream is removed from the received signal. The second term, in turn, can be split into a rate achieved by MMSE detection of the second stream and a rate corresponding to all other streams. Proceeding recursively in this way, we obtain,

$$I(\mathsf{s}_{k,1},\ldots,\mathsf{s}_{k,m_k};\boldsymbol{r}_k^{(1)}) = \sum_{j=1}^{K} I(\mathsf{s}_{k,j};\boldsymbol{r}_k^{(j)}) = \sum_{j=1}^{K} I(\mathsf{s}_{k,j};\hat{\mathsf{s}}_{k,j}),$$

where $\boldsymbol{r}_k^{(j)} = \sum_{i=j}^{m_k} \boldsymbol{f}_{k,i}\mathsf{s}_{k,i}$ and $\hat{\mathsf{s}}_{k,j}$ is the MMSE estimate of $s_{k,j}$ based on the observation $\boldsymbol{r}_k^{(j)}$. That is, for the given input statistics, the maximum achievable rate for user k can be expanded as a sum of rates corresponding to each of the individual streams. The rate corresponding to a particular stream can be achieved by applying an MMSE estimator to the received signal without the contributions of previously detected streams and decoding information based on this estimate. Note that in this exposition the streams have been decoded in the order indicated by their indexes. This has been done in order to simplify notation. Choosing a different ordering would change the rates corresponding to the particular streams but not the total rate resulting from the addition of the individual stream rates.

Thus, we conclude that, at the transmitter, the transmit signal can be decomposed into a set of statistically independent streams and, at the receiver, these streams can be independently decoded, using the outputs of an MMSE-DFE, without loss of optimality. That is, given any input statistics for the MAC, there is always a streamwise transmission strategy that achieves the rates that are achievable with those statistics. In the context of time-dispersive single-input single-output channels, optimality of the MMSE-DFE in terms of capacity was first shown in [32]. This result was also shown in [126] applied to multiple access channels.

A.4.2 Streamwise duality

Consider a streamwise strategy in the MAC for which the transmit signals are given by

$$\boldsymbol{w}_k = \sum_{j=1}^{m_k} \boldsymbol{u}_{k,j} q_{k,j}^{1/2} \mathsf{s}_{k,j}, \quad k = 1,\ldots,K.$$

Here, $\boldsymbol{u}_{k,j}$ is a unit-norm beamforming vector corresponding to the stream j of user k and $q_{k,j}$ is the power assigned to this stream. As in the previous section, the signals $\mathsf{s}_{k,j}$, $k = 1,\ldots,K$, $j = 1,\ldots,m_k$, are statistically independent and Gaussian with zero mean and unit variance. Assume that the streams are successively decoded in the order indicated by the user index and the stream index, i.e., first, the first stream of user 1 is decoded, second, the second stream of this user, then, after all streams of user 1 have been decoded, the first stream of user 2 is decoded, and so on. Correspondingly, before stream $s_{k,j}$ is decoded, interference caused by the streams of users $1,\ldots,k-1$ and the streams $1,\ldots,j-1$ of user k can be removed from the received signal. On the resulting effective received signal a unit-norm linear filter $\boldsymbol{v}_{k,j}$ is applied and detection of $s_{k,j}$ is done based on the output

$\hat{s}_{k,j}$ of this filter, which reads

$$\hat{s}_{k,j} = \boldsymbol{v}_{k,j}^{\mathrm{H}} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{u}_{k,j} q_{k,j}^{1/2} s_{k,j} + \boldsymbol{v}_{k,j}^{\mathrm{H}} \boldsymbol{H}_k^{\mathrm{H}} \sum_{\ell=j+1}^{m_k} \boldsymbol{u}_{k,\ell} q_{k,\ell}^{1/2} s_{k,\ell} + \boldsymbol{v}_{k,j}^{\mathrm{H}} \sum_{i=k+1}^{K} \sum_{\ell=1}^{m_k} \boldsymbol{H}_i^{\mathrm{H}} \boldsymbol{u}_{i,\ell} q_{i,\ell}^{1/2} s_{i,\ell} + \boldsymbol{v}_{k,j}^{\mathrm{H}} \boldsymbol{z}.$$

The maximum rate achievable on this stream is given by $R_{k,j} = \log_2(1 + \mathrm{SINR}_{k,j}^{\mathrm{MAC}})$, where the signal-to-interference-plus-noise ratio $\mathrm{SINR}_{k,j}^{\mathrm{MAC}}$ is defined as

$$\mathrm{SINR}_{k,j}^{\mathrm{MAC}} = \frac{q_{k,j} |\boldsymbol{v}_{k,j}^{\mathrm{H}} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{u}_{k,j}|^2}{1 + \sum_{\ell=j+1}^{m_k} |\boldsymbol{v}_{k,j}^{\mathrm{H}} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{u}_{k,\ell}|^2 q_{k,\ell} + \sum_{i=k+1}^{K} \sum_{\ell=1}^{m_k} |\boldsymbol{v}_{k,j}^{\mathrm{H}} \boldsymbol{H}_i^{\mathrm{H}} \boldsymbol{u}_{i,\ell}|^2 q_{i,\ell}}.$$

The following dual streamwise strategy could be employed for transmission over the dual BC. The transmitted signal is given by

$$\boldsymbol{x} = \sum_{k=1}^{K} \sum_{j=1}^{m_k} \boldsymbol{v}_{k,j} p_{k,j}^{1/2} s_{k,j},$$

where $p_{k,j}$ is the power assigned to the jth stream intended for user k. These streams are successively encoded using a dirty paper coding strategy in order to neutralize interference caused by previously encoded streams (cf. Section 2.2.2.1). Assume that the encoding order is the reverse of the decoding order assumed above, i.e., the stream m_K of user K is first encoded, second, stream $m_K - 1$ of this user is encoded, then, once all streams of user K have been encoded, the stream m_{K-1} of user $K-1$ is encoded and so on. At the receivers, detection of a specific stream is based on the output of a linear filter. In particular, for the jth stream of user k we have

$$\hat{s}_{k,j} = \boldsymbol{u}_{k,j}^{\mathrm{H}} \boldsymbol{H}_k \boldsymbol{v}_{k,j} p_{k,j}^{1/2} s_{k,j} + \boldsymbol{u}_{k,j}^{\mathrm{H}} \boldsymbol{H}_k \sum_{\ell=1}^{j-1} \boldsymbol{v}_{k,\ell} p_{k,\ell}^{1/2} s_{k,\ell} + \boldsymbol{u}_{k,j}^{\mathrm{H}} \boldsymbol{H}_k \sum_{i=1}^{k-1} \sum_{\ell=1}^{m_k} \boldsymbol{v}_{i,\ell} p_{i,\ell}^{1/2} s_{i,\ell} + \boldsymbol{u}_{k,j}^{\mathrm{H}} \boldsymbol{n}.$$

As in the MAC, here also, the rate achievable on a stream is characterized by the signal-to-interference-plus-noise ratio of that stream defined as

$$\mathrm{SINR}_{k,j}^{\mathrm{BC}} = \frac{p_{k,j} |\boldsymbol{v}_{k,j}^{\mathrm{H}} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{u}_{k,j}|^2}{1 + \sum_{\ell=1}^{j-1} |\boldsymbol{v}_{k,\ell}^{\mathrm{H}} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{u}_{k,j}|^2 p_{k,\ell} + \sum_{i=1}^{k-1} \sum_{\ell=1}^{m_k} |\boldsymbol{v}_{i,\ell}^{\mathrm{H}} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{u}_{k,j}|^2 p_{i,\ell}}.$$

In order to show that the rates achievable on the streams of the MAC are also achievable on the streams of the BC, or vice versa, we proceed along the lines of the proof of the MMSE-based BC-MAC duality given in [82]. Assume that the stream powers $q_{k,j}$, $k = 1, \ldots, K$, $j = 1, \ldots, m_k$ are given that achieve certain SINR values in the MAC. We want to find out whether there is a set of powers $p_{k,j}$, $k = 1, \ldots, K$, $j = 1, \ldots, m_k$, that achieve the same SINR values for the streams in the BC and such that $\sum_{k=1}^{K} \sum_{j=1}^{m_k} p_{k,j} = \sum_{k=1}^{K} \sum_{j=1}^{m_k} q_{k,j}$. To this end, consider the linear system of $M = \sum_{k=1}^{K} m_k$ equations with the M unknowns $p_{k,j}$, $k = 1, \ldots, K$, $j = 1, \ldots, m_k$, that is obtained by setting $\mathrm{SINR}_{k,j}^{\mathrm{BC}} = \mathrm{SINR}_{k,j}^{\mathrm{MAC}}$, $k = 1, \ldots, K$, $j = 1, \ldots, m_k$. That is,

$$\boldsymbol{A}\boldsymbol{p} = \boldsymbol{q}, \tag{A.12}$$

where $\boldsymbol{p} = [\ \boldsymbol{p}_1^{\mathrm{T}} \ \cdots \ \boldsymbol{p}_K^{\mathrm{T}}\ ]^{\mathrm{T}}$, $\boldsymbol{p}_k = [\ p_{k,1} \ \cdots \ p_{k,m_k}\ ]^{\mathrm{T}}$, $k = 1, \ldots, K$, and $\boldsymbol{q}$ is defined similarly. Matrix $\boldsymbol{A}$ is an $M \times M$ lower triangular matrix. The column $\sum_{i=1}^{k-1} m_k + j$ of this matrix is given by

$$\boldsymbol{a}_{k,j} = [\ 0 \ \cdots \ 0 \ \ d_{k,j} \ \ \boldsymbol{i}_k^{\mathrm{T}} \ \ \boldsymbol{i}_{k+1}^{\mathrm{T}} \ \cdots \ \boldsymbol{i}_K^{\mathrm{T}}\ ]^{\mathrm{T}},$$

where

$$d_{k,j} = 1 + \sum_{\ell=j+1}^{m_k} |\boldsymbol{v}_{k,j}^{\mathrm{H}} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{u}_{k,\ell}|^2 q_{k,\ell} + \sum_{i=k+1}^{K} \sum_{\ell=1}^{m_k} |\boldsymbol{v}_{k,j}^{\mathrm{H}} \boldsymbol{H}_i^{\mathrm{H}} \boldsymbol{u}_{i,\ell}|^2 q_{i,\ell}$$

is the entry on the main diagonal,

$$\boldsymbol{i}_k = [\ -q_{k,j+1} |\boldsymbol{v}_{k,j}^{\mathrm{H}} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{u}_{k,j+1}|^2 \ \cdots \ -q_{k,m_k} |\boldsymbol{v}_{k,j}^{\mathrm{H}} \boldsymbol{H}_k^{\mathrm{H}} \boldsymbol{u}_{k,m_k}|^2\ ]^{\mathrm{T}}$$

and

$$\boldsymbol{i}_i = [\ -q_{i,1} |\boldsymbol{v}_{k,j}^{\mathrm{H}} \boldsymbol{H}_i^{\mathrm{H}} \boldsymbol{u}_{i,1}|^2 \ \cdots \ -q_{i,m_i} |\boldsymbol{v}_{k,j}^{\mathrm{H}} \boldsymbol{H}_i^{\mathrm{H}} \boldsymbol{u}_{i,m_i}|^2\ ]^{\mathrm{T}}, \quad i = k+1, \ldots, K.$$

Using the fact that $d_{k,j} > \sum_{i=k}^{K} \|\boldsymbol{i}_i\|_1$ and that all off-diagonal entries are negative, it is straightforward to show that the inverse of $\boldsymbol{A}$ has positive entries. This ensures the existence of a power vector $\boldsymbol{p}$ that achieves in the BC the same performance as that achieved by $\boldsymbol{q}$ in the MAC. Furthermore, since for every column of matrix $\boldsymbol{A}$ we have $\mathbf{1}^{\mathrm{T}} \boldsymbol{a}_{k,j} = 1$, multiplying the left- and the right-hand sides of Eq. A.12 by $\mathbf{1}^{\mathrm{T}}$, $\|\boldsymbol{p}\|_1 = \|\boldsymbol{q}\|_1$ immediately follows. Similar duality results have been shown in [122, 132, 100] resorting to Perron-Frobenius theory.

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, 1964.

[2] R. Ahlswede. Multi-way communication channels. In *Proceedings of 2nd International Symposium on Information Theory*, 1971.

[3] D. Astely, E. Dahlman, P. Frenger, R. Ludwig, M. Meyer, S. Parkvall, P. Slillermark, and N. Wiberg. A future radio-access framework. *IEEE Journal on Selected Areas in Communications*, Vol. 24:693–706, Mar. 2006.

[4] M. S. Bazaara, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming: theory and algorithms*. Wiley-Interscience, 2006.

[5] P. P. Bergmans. Random coding theorem for broadcast channels with degraded components. *IEEE Trans. on Information Theory*, Vol. 19:197–207, Mar. 1973.

[6] C. Berrou, A. Glavieux, and P. Thitimajshima. Near Shannon limit error-correcting coding and decoding: turbo-codes. In *IEEE International Conference on Communications*, 2003.

[7] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[8] H. Boche and M. Wiczanowski. Optimal scheduling for high speed uplink packet access - a cross-layer approach. In *IEEE Vehicular Technology Conference*, May 2004.

[9] H. Boche and M. Wiczanowski. Optimization-theoretic analysis of stability-optimal transmission policy for multiple-antenna multiple-access channel. *IEEE Trans. on Signal Processing*, 55:2688–2702, Jun. 2007.

[10] R. Böhnke and K. D. Kammeyer. Weighted sum rate maximization for the MIMO-downlink using a projected conjugate gradient algorithm. In *First International Workshop on Cross Layer Design*, 2007.

[11] R. Böhnke, V. Kühn, and K. D. Kammeyer. Fast sum rate maximization for the downlink of MIMO-OFDM systems. In *Canadian Workshop on Information Theory*, 2005.

[12] S. Boyd. Convex Optimization II, Course material. Available at www.stanford.edu/class/ee364b/.

[13] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambrige University Press, 2004.

[14] S. Boyd, L. Xiao, and A. Mutapcic. Subgradient methods. Notes for EE392o, Standford University, 2003.

[15] J. Brehmer, Q. Bai, and W. Utschick. Time-sharing solutions in MIMO broadcast channel utility maximization. In *ITG Workshop on Smart Antennas*, 2008.

[16] J. Brehmer, A. Molin, P. Tejera, and W. Utschick. Low complexity approximation of the MIMO broadcast channel capacity region. In *IEEE International Conference on Communications*, 2006.

[17] X. Cai and J. W. Modestino. Bandwidth expansion shannon mapping for analog error-control coding. In *40th Annual Conference on Information Sciences and Systems (CISS)*, 2006.

[18] G. Caire. MIMO downlink joint processing and scheduling: a survey of classical and recent results. In *Workshop on Information Theory and its Applications*, 2006.

[19] G. Caire, N. Jindal, and M. Kobayashi. Achievable rates of MIMO downlink beamforming with non-perfect CSI: a comparison between quantized and analog feedback. In *Asilomar Conference on Signals, Systems and Computers*, 2006.

[20] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran. Quantized vs. analog feedback for the MIMO broadcast channel: A comparison between zero-forcing based achievable rates. In *IEEE International Symposium on Information Theory (ISIT)*, 2007.

[21] G. Caire and S. Shamai. On achievable rates in a multi-antenna broadcast downlink. In *Proc. 38th Annual Allerton Conf. Communication, Control and Computing.*, 2000.

[22] G. Caire and S. Shamai. On achievable rates in a multi-antenna Gaussian broadcast channel. In *IEEE International Symposium on Information Theory (ISIT)*, 2001.

[23] G. Caire and S. Shamai. On the achievable throughput of a multi-antenna Gaussian broadcast channel. *IEEE Trans. on Information Theory*, Vol. 49:1691–1706, Jul. 2003.

[24] G. Caire and S. Shamai. On the capacity of some channels with channel state information. *IEEE Trans. on Information Theory*, Vol. 45:2007–2019, Sep. 1999.

[25] G. Caire, G. Taricco, and E. Biglieri. Bit-interleaved coded modulation. *IEEE Trans. on Information Theory*, Vol. 44:927–946, May 1998.

[26] R. Cendrillon, G. Ginis, E. van den Bogaert, and M. Moonen. A near-optimal linear crosstalk precoder for downstream VDSL. *IEEE Trans. on Communications*, Vol. 55:860–863, May 2007.

[27] R. Cendrillon, M. Moonen, R. Suciu, and G. Ginis. Simplified power allocation and TX/RX structure for MIMO-DSL. In *IEEE Global Telecommunications Conference*, 2002.

[28] B. Chen and G. W. Wornell. Analog error-correcting codes based on chaotic dynamical systems. *IEEE Trans. on Communications*, Vol. 46:881–890, Jul. 1998.

[29] J. Choi, B. Mondal, and R. W. Heath. Interpolation based unitary precoding for spatial multiplexing MIMO-OFDM with limited feedback. *IEEE Trans. on Signal Processing*, Vol. 54:4730–4740, Dec. 2006.

[30] L. Choi, M. T. Ivrlac, R. D. Murch, and J. A. Nossek. Joint transmit and receive multi-user MIMO decomposition approach for the downlink of multi-user MIMO systems. In *IEEE Vehicular Technology Conference*, Oct. 2003.

[31] L. Choi and R. D. Murch. A transmit preprocessing technique for multiuser MIMO systems using a decomposition approach. *IEEE Trans. on Wireless Communication*, Vol. 3:20–24, Jan. 2004.

[32] J. M. Cioffi, G. P. Dudevoir, M. V. Eyuboglu, and G. D. Forney. MMSE decision-feedback equalizers and coding. II. coding results. *IEEE Trans. on Communications*, Vol. 43:2595 – 2604, Oct. 1995.

[33] M. Codreanu, M. Juntti, and M. Latva-Aho. Low-complexity iterative algorithm for finding the MIMO-OFDM broadcast channel sum capacity. *IEEE Transactions on Communications*, Vol. 55:48–53, Jan. 2007.

[34] M. Codreanu, M. Juntti, and M. Latva-Aho. On the dual decomposition based sum capacity maximization for vector broadcast channels. *IEEE Transactions on Vehicular Technology*, Vol. 56:3577–3581, Nov. 2007.

[35] A. S. Cohen and A. Lapidoth. Generalized writing on dirty paper. In *IEEE International Symposium on Information Theory (ISIT)*, 2002.

[36] J. H. Conway and N. J. A. Sloane. *Sphere Packings, Lattices and Groups*. Springer-Verlag, 1999.

[37] M. Costa. Writing on dirty paper. *IEEE Trans. on Information Theory*, Vol. 29:439–441, May 1983.

[38] T. M. Cover. Broadcast channels. *IEEE Trans. on Information Theory*, Vol. 18:2–14, Jan. 1972.

[39] T. M. Cover. Comments on broadcast channels. *IEEE Trans. on Information Theory*, Vol. 44:2524–2530, May 1998.

[40] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 2005.

[41] A. D. Dabbagh and D. J. Love. Precoding for multiple antenna Gaussian broadcast channels with successive zero-forcing. *IEEE Trans. on Signal Processing*, Vol. 55:3837–3850, Jul. 2007.

[42] G. Dimić and N. D. Sidiropoulos. On downlink beamforming with greedy user selection: Performance analysis and a simple new algorithm. *IEEE Trans. on Signal Processing*, Vol. 53:3857–3868, Oct. 2005.

[43] P. Ding, D. J. Love, and M. D. Zoltowski. Multiple antennas broadcast channels with shape feedback and limited feedback. *IEEE Trans. on Signal Processing*, Vol. 55:3417 – 3428, Jul. 2007.

[44] H. Ekström, A. Furuskär, J. Karlsson, M. Meyer, S. Parkvall, J. Torsner, and M. Wahlqvist. Technical solutions for the 3G long-term evolution. *IEEE Communications Magazine*, Vol. 44:38–48, Mar. 2006.

[45] A. El Gamal and E. C. van der Meulen. A proof of Marton's coding theorem for the discrete memoryless Broadcast Channel. *IEEE Trans. on Information Theory*, Vol. 27:120–122, Jan. 1981.

[46] U. Erez, S. Shamai, and R. Zamir. Capacity and lattice strategies for cancelling known interference. *IEEE Trans. on Information Theory*, Vol. 51:3820 – 3833, Nov. 2005.

[47] N. Farvardin and V. Vaishampayan. Optimal quantizer design for noisy channels: An approach to combined source-channel coding. *IEEE Trans. on Information Theory*, Vol. 33:827–838, Nov. 1987.

[48] B. Friedrichs. *Kanalcodierung: Grundlagen und Anwendungen in modernen Kommunikationssystemen*. Springer Verlag, 1995.

[49] M. Fuchs, G. D. Galdo, and M. Haardt. Low-complexity space-time-frequency scheduling for MIMO systems with SDMA. *IEEE Trans. on Vehicular Technology*, Vol. 56:2775–2784, Sep. 2007.

[50] C.-H. F. Fung, W. Yu, and T. J. Lim. Multi-antenna downlink precoding with individual rate constraints: power minimization and user ordering. In *Conference on Information Sciences and Systems*, 2004.

[51] N. T. Gaarder and D. Slepian. On optimal finite-state digital transmission systems. *IEEE Trans. on Information Theory*, Vol. 28:167–186, Mar. 1982.

[52] R. G. Gallager. *Information theory and reliable communication*. John Wiley & Sons, 1968.

[53] M. Gastpar, B. Rimoldi, and M. Vetterli. To code or not to code: lossy source-channel communication revisited. *IEEE Trans. on Information Theory*, Vol. 49:1147–1158, May 2003.

[54] S. Gelfand and M. Pinsker. Coding for channel with random paramenters. *Problems of Control and Information Theory*, Vol. 9:19–31, Jan. 1980.

[55] A. Gersho and R. M. Gray. *Vector quantization and signal compression*. Kluwer Academic Publishers, 1991.

[56] G. Ginis and J. Cioffi. A multi-user precoding scheme achieving crosstalk cancellation with application to DSL systems. In *Asilomar Conference on Signals, Systems and Computers*, pages 1627–1631, 2000.

[57] T. J. Goblick. Theoretical limitations on the transmission of data from analog sources. *IEEE Trans. on Information Theory*, Vol. 11:558–567, Oct. 1965.

[58] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The John Hopkins University Press, 1989.

[59] F. Hekland. A review of joint source-channel coding. available at www.iet.ntnu.no/ hekland/, 2004.

[60] B. Hochwald and K. Zeger. Tradeoff between source and channel coding. *IEEE Trans. on Information Theory*, Vol. 43:1412–1424, Sep. 1997.

[61] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambrige University Press, 1991.

[62] R. Hunger, D. Schmidt, M. Joham, and W. Utschick. A general covariance-based optimization framework using orthogonal projections. In *Proceedings of the International Workshop on Signal Processing Advances in Wireless Communications*, 2008.

[63] R. Hunger, D. Schmidt, and W. Utschick. Sum-capacity and MMSE for the MIMO broadcast channel without eigenvalue decompositions. In *IEEE International Symposium on Information Theory (ISIT)*, 2007.

[64] N. Jindal. MIMO broadcast channels with finite rate feedback. *IEEE Trans. on Information Theory*, Vol. 52:5045–5060, Nov. 2006.

[65] N. Jindal, W. Rhee, S. Vishwanath, S. A. Jafar, and A. Goldsmith. Sum power iterative water-filling for multi-antenna Gaussian broadcast channels. *IEEE Trans. on Information Theory*, Vol. 51:1570–1580, Feb. 2005.

[66] G. Kaplan and S. Shamai. Error probabilities for the block-fading Gaussian channel. *Archiv für Elektronik und Übertragungstechnik*, Vol. 49:192–205, 1995.

[67] S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, Inc., 1993.

[68] M. Kobayashi and G. Caire. Iterative waterfilling for weighted sum rate maximization in MIMO-OFDM broadcast channels. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[69] M. Kobayashi and G. Caire. An iterative water-filling algorithm for maximum weighted sum-rate of Gaussian MIMO-BC. *IEEE Journal on Selected Areas in Communications*, Vol. 24:1640–1646, Aug. 2006.

[70] A. J. Kurtenbach and P. A. Wintz. Quantizing for noisy channels. *Trans. on Information Theory*, COM-17:291–302, Apr. 1969.

[71] J. Lee and N. Jindal. Symmetric capacity of MIMO downlink channels. In *IEEE Internatinal Symposium on Information Theory (ISIT)*, 2006.

[72] T. Linder and R. Zamir. Causal coding of stationary sources and individual sequences with high resolution. *IEEE Trans. on Information Theory*, Vol. 52:662–680, Feb. 2006.

[73] F. H. Liu, P. Ho, and V. Cuperman. Joint source and channel coding using a non-linear receiver. In *IEEE International Conference on Communications*, 1993.

[74] F. H. Liu, P. Ho, and V. Cuperman. Joint source and channel coding using a non-linear receiver over Rayleigh fading channel. In *IEEE Global Telecommunications Conference*, 1993.

[75] J. Liu and Y. T. Hou. Maximum weighted sum rate of multi-antenna broadcast channels. In *IEEE Global Telecommunications Conference*, 2007.

[76] T. D. Lookabaugh and R. M. Gray. High-resolution quantization theory and the vector quantizer advantage. *IEEE Trans. on Information Theory*, Vol. 35:1020 – 1033, Sep. 1989.

[77] D. J. Love, R. W. Heath, W. Santipach, and M. Honig. What is the value of limited feedback for MIMO channels. *IEEE Communications Magazine*, Vol. 42:54–59, Oct. 2004.

[78] D. J. Love, R. W. Heath, and T. Strohmer. Grassmannian beamforming for multiple-input multiple-output wireless systems. *IEEE Trans. on Information Theory*, Vol. 49:2735–2747, Oct. 2003.

[79] M. A. Maddah-Ali and A. K. K. M. Ansari. An efficient signaling method over MIMO broadcast channels. In *Proc. 42th Annual Allerton Conf. Communication, Control and Computing.*, 2004.

[80] K. Marton. A coding theorem for the discrete memoryless broadcast Channel. *IEEE Trans. on Information Theory*, Vol. 25:306–311, May 1979.

[81] T. L. Marzetta and B. M. Hochwald. Fast transfer of channel state information in wireless systems. *IEEE Trans. on Signal Processing*, Vol. 54:1268–1278, Apr. 2006.

[82] A. Mezghani, M. Joham, R. Hunger, and W. Utschick. Iterative THP transceiver optimization for multi-user MIMO systems based on weighted sum-MSE minimization. In *International Workshop on Signal Processing Advances in Wireless Communications*, 2006.

[83] T. Michel and G. Wunder. Optimal and low complex suboptimal transmission schemes for MIMO-OFDM broadcast channels. In *IEEE International Conference on Communications*, 2005.

[84] K. K. Mukkavilli, A. Sabharwal, E. Erkip, and A. Aazhang. On beamforming with finite rate feedback in multiple-antenna systems. *IEEE Trans. on Information Theory*, Vol. 49:2562–2579, Oct. 2003.

[85] A. Nedić and A. Ozdaglar. Approximate primal solutions and rate analysis for dual subgradient methods. LIDS technical report 2753, Massachusetts Institute of Technology, Lab. for Information and Decision Systems, March 2007.

[86] A. Nemirovski. Advances in convex optimization: Conic programming. In *Plenary lecture at the International Congress of Mathematicians*, Madrid, August 2006.

[87] F. Oberhettinger and L. Badii. *Tables of Laplace Transforms*. Springer Verlag, 1973.

[88] T. Pande, D. J. Love, and J. V. Krogmeier. Reduced feedback MIMO-OFDM precoding and antenna selection. *IEEE Trans. on Signal Processing*, Vol. 55:2284–2293, May 2007.

[89] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. Mc-Graw-Hill, Inc., 1991.

[90] D. Perez Palomar and J. Rodriguez Fonollosa. Practical algorithms for a family of waterfilling solutions. *IEEE Transactions on Signal Processing*, Vol. 53:686–695, Feb. 2005.

[91] N. Ravindran and N. Jindal. MIMO broadcast channels with block diagonalization and finite rate feedback. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[92] B. Riedmüller and K. Ritter. *Lineare und quadratische Optimierung*. Institut für Angewandte Mathematik und Statistik. Technische Universität München, 1992.

[93] M. Rim. Multiuser downlink beamforming with multiple transmit and receive Antennas. *Electronics Letters*, Vol. 38:1725–1726, Dec. 2002.

[94] K. Ritter. *Nichtlineare Optimierung*. Institut für Angewandte Mathematik und Statistik. Technische Universität München, 1992.

[95] J. C. Roh and B. D. Rao. Design and analysis of MIMO spatial multiplexing systems with quantized feedback. *IEEE Trans. on Signal Processing*, Vol. 54:2874–2886, Aug. 2006.

[96] D. Samardzija and N. Mandayam. Unquantized and uncoded channel state information feedback in mulitple-antenna multiuser systems. *IEEE Trans. on Communications*, Vol. 54:1335–1345, Jul. 2006.

[97] D. Samardzija and N. Mandayam. Multiple antenna transmitter optimization schemes for multiuser systems. In *IEEE Vehicular Technology Conference*, Oct. 2003.

[98] H. Sampath, S. Talwar, J. Tellado, V. Erceg, and A. Paulraj. A fourth-generation MIMO-OFDM broadband wireless system: design, performance, and field trial results. *IEEE Communications Magazine*, Vol. 40:143–149, Sep. 2002.

[99] H. Sato. An outer bound on the capacity region of broadcast channels. *IEEE Trans. on Information Theory*, Vol. 24:374–377, May 1978.

[100] M. Schubert and H. Boche. Iterative multiuser uplink and downlink beamforming under SINR constraints. *IEEE Trans. on Signal Processing*, Vol. 53:2324–2334, Jul. 2005.

[101] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, Vol. 27:379–423, 623–656, 1948.

[102] C. E. Shannon. Communication in the presence of noise. *IRE*, Vol. 37:10–21, 1949.

[103] Z. Shen, J. G. Andrews, and B. L. Evans. Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints. *IEEE Trans. on Wireless Communications*, Vol. 4:2726–2737, Nov. 2005.

[104] Z. Shen, R. Chen, J. G. Andrews, R. W. Heath, and B. L. Evans. Low complexity user selection algorithms for multiuser MIMO systems with block diagonalization. *IEEE Trans. on Signal Processing*, Vol. 54:3658 – 3663, Sep. 2006.

[105] H. Shin and J. H. Lee. Capacity of multiple-antenna fading channels: Spatial fading correlation, double scattering, and keyhole. *IEEE Trans. on Information Theory*, Vol. 49:2636 – 2647, Oct. 2003.

[106] Q. H. Spencer and A. L. Swindlehurst. Channel allocation in multi-user MIMO wireless communications systems. In *IEEE International Conference on Communications*, 2004.

[107] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt. Zero forcing methods for downlink spatial multiplexing in multiuser MIMO channels. *IEEE Trans. on Signal Processing*, Vol. 52:461 – 470, Feb. 2004.

[108] V. Stankovic and M. Haardt. Successive optimization Tomlinson-Harashima precoding (SO-THP) for multi-user MIMO systems. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2005.

[109] K. Tachikawa. A perspective on the evolution of mobile communications. *IEEE Communications Magazine*, Vol. 41:66–73, Oct. 2003.

[110] P. Tejera, C. K. Schmidt, and W. Utschick. Rate balancing in the broadcast channel with independent encoding. In *ITG/IEEE Workshop on Smart Antennas*, 2007.

[111] P. Tejera and W. Utschick. Feedback of channel state information in wireless systems. In *IEEE International Conference on Communications*, 2007.

[112] P. Tejera and W. Utschick. Delay-limited feedback of channel state information in OFDM systems. In *7th International Symposium on Source and Channel Coding*, 2008.

[113] P. Tejera, W. Utschick, G. Bauch, and J. A. Nossek. A novel decomposition technique for multiuser MIMO. In *IEEE/ITG Workshop on Smart Antennas*, 2005.

[114] P. Tejera, W. Utschick, G. Bauch, and J. A. Nossek. Analysis of the impact of channel estimation errors on the decomposition of multiuser MIMO channels. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006.

[115] P. Tejera, W. Utschick, G. Bauch, and J. A. Nossek. Efficient implementation of successive encoding schemes for the MIMO OFDM broadcast channel. In *IEEE International Conference on Communications*, 2006.

[116] P. Tejera, W. Utschick, G. Bauch, and J. A. Nossek. Rate balancing in multiuser MIMO OFDM systems. *accepted for publication in IEEE Transactions on Communications*, 2008.

[117] P. Tejera, W. Utschick, G. Bauch, and J. A. Nossek. Subchannel allocation in multiuser multiple input multiple output systems. *IEEE Trans. on Information Theory*, Vol. 52:4721–4732, Oct. 2006.

[118] P. Tejera, W. Utschick, G. Bauch, and J. A. Nossek. Sum-rate maximizing decomposition approaches for multiuser MIMO-OFDM. In *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Berlin, Sep. 2005.

[119] E. Telatar. Capacity of multi-antenna Gaussian channels. *European Transactions on Telecommunications*, Vol. 10(6):585–595, 1999.

[120] T. A. Thomas, K. L. Baum, and P. Sartori. Obtaining channel knowledge for closed-loop multi-stream broadband MIMO-OFDM communications using direct channel feedback. In *IEEE Global Telecommunications Conference*, 2005.

[121] D. N. C. Tse and S. V. Hanly. Multiaccess fading channels - Part I: Polymatroid structure, optimal resource allocation and throughput capacities. *IEEE Trans. on Information Theory*, Vol. 44:2796–2815, 1998.

[122] D. N. C. Tse and P. Viswanath. On the capacity of the multiple antenna broadcast channel. In G. J. Foschini and S. Verdú, editors, *Multiantenna Channels: Capacity, Coding and Signal Processing*. American Mathematical Society, DIMACS, 2003.

[123] Z. Tu and R. S. Blum. Multiuser diversity for a dirty paper approach. *IEEE Communications Letters*, Vol. 7:370–372, Aug. 2003.

[124] M. Uppal, V. Stankovic, and Z. Xiong. Code designs for MIMO broadcast channels. In *IEEE International Symposium on Information Theory (ISIT)*, 2006.

[125] V. A. Vaishampayan and N. Farvardin. Joint design of block source codes and modulation signal sets. *IEEE Trans. on Information Theory*, Vol. 38:1230–1248, Jul. 1992.

[126] M. K. Varanasi and T. Guess. Optimum decision feedback multiuser equalization with successive decoding achieves the total capacity of the Gaussian multiple-access channel. In *Asilomar Conference on Signals, Systems and Computers*, 1997.

[127] S. Vishwanath, N. Jindal, and A. Goldsmith. On the capacity of multiple input multiple output broadcast channels. In *IEEE International Conference on Communications (ICC)*, 2002.

[128] S. Vishwanath, N. Jindal, and A. Goldsmith. Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels. *IEEE Trans. on Information Theory*, Vol. 49:2658–2668, Oct. 2003.

[129] S. Vishwanath, G. Kramer, S. Shamai, S. Jafar, and A. Goldsmith. Capacity bounds for Gaussian vector broadcast channels. In G. J. Foschini and S. Verdú, editors, *Multiantenna Channels: Capacity, Coding and Signal Processing*. American Mathematical Society, DIMACS, 2003.

[130] S. Vishwanath, W. Rhee, N. Jindal, S. Jafar, and A. Goldsmith. Sum power iterative waterfilling for Gaussian vector broadcast channels. In *IEEE International Symposium on Information Theory (ISIT)*, 2003.

[131] E. Visotsky and U. Madhow. Space-time transmit strategies and channel feedback generation for wireless fading channels. In *Asilomar Conference on Signals, Systems and Computers*, 2001.

[132] P. Viswanath and D. N. C. Tse. Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality. *IEEE Trans. on Information Theory*, Vol. 49:1912 – 1921, Aug. 2003.

[133] H. Viswanathan, S. Venkatesan, and H. Huang. Downlink capacity evaluation of cellular networks with known interference cancellation. *IEEE Journal on Selected Areas in Communications*, Vol. 21:802–811, Jun. 2003.

[134] H. Weingarten, Y. Steinberg, and S. Shamai. The capacity region of the Gaussian MIMO broadcast channel. In *Conference on Information Sciences and Systems*, Princeton University, 2004.

[135] H. Weingarten, Y. Steinberg, and S. Shamai. The capacity region of the Gaussian multiple-input multiple-output broadcast channel. *IEEE Trans. on Information Theory*, Vol. 52:3936–3964, Sep. 2006.

[136] C. Windpassinger, T. Vencel, and R. F. H. Fischer. Precoding and loading for BLAST-like systems. In *IEEE International Conference on Communications*, pages 3061–3065, Anchorage, Alaska, 2003.

[137] Y. Wu, J. Zhang, H. Zheng, X. Xu, and S. Zhou. Receive antenna selection in the downlink of multiuser MIMO systems. In *IEEE Vehicular Technology Conference*, Sep. 2005.

[138] H. Yang. A road to future broadband wireless access: MIMO-OFDM-based air interface. *IEEE Communications Magazine*, Vol. 43:53–60, Jan. 2005.

[139] T. Yoo and A. Goldsmith. On the optimality of multi-antenna broadcast scheduling using zero-forcing beamforming. *IEEE Journal on Selected Areas in Communications*, Vol. 24:528–541, Mar. 2006.

[140] W. Yu. Dirty paper: the vector case. Available at citeseer.ist.psu.edu/yu00dirty.html.

[141] W. Yu. A dual decomposition approach to the sum power Gaussian vector multiple-access channel sum capacity problem. In *Conference on Information Sciences and Systems*, 2003.

[142] W. Yu. Sum-capacity computation for the Gaussian vector broadcast channel via dual decomposition. *IEEE Trans. on Information Theory*, Vol. 52:754 – 759, Feb. 2006.

[143] W. Yu and J. M. Cioffi. Sum capacity of a Gaussian vector broadcast channel. In *IEEE International Symposium on Information Theory (ISIT)*, 2002.

[144] W. Yu and J. M. Cioffi. Sum capacity of Gaussian vector broadcast channels. *IEEE Trans. on Information Theory*, Vol. 50:1875 – 1892, Sep. 2004.

[145] W. Yu and J. M. Cioffi. Trellis precoding for the broadcast channel. In *IEEE Global Telecommunications Conference*, 2001.

[146] W. Yu, W. Rhee, S. Boyd, and J. M. Cioffi. Iterative water-filling for Gaussian vector multiple-access channels. In *IEEE International Symposium on Information Theory (ISIT)*, 2001.

[147] W. Yu, W. Rhee, S. Boyd, and J. M. Cioffi. Iterative water-filling for Gaussian vector multiple-access channels. *IEEE Trans. on Information Theory*, Vol. 50:145 – 152, Jan. 2004.

[148] W. Yu, A. Sutivong, D. Julian, T. M. Cover, and M. Chiang. Writing on colored paper. In *IEEE International Symposium on Information Theory (ISIT)*, 2001.

[149] W. Yu, D. P. Varodayan, and J. M. Cioffi. Trellis and convolutional precoding for transmitter-based presubtraction. *IEEE Trans. on Communications*, Vol. 53:1220–1230, Jul. 2005.

[150] S. B. Zahir Azami, P. Duhamel, and O. Rioul. Combined source-channel coding: Panorama of methods. In *CNES Workshop on Data Compression*, 1996.

[151] R. Zamir, S. Shamai, and U. Erez. Nested linear/lattice codes for structured multiterminal binning. *IEEE Trans. on Information Theory*, Vol. 48:1250–1276, Jun. 2002.

[152] K. Zeger and A. Gersho. Pseudo-Gray coding. *IEEE Trans. on Communications*, Vol. 38:2147–2158, Dec. 1990.

[153] K. A. Zeger and A. Gersho. Vector quantizer design for memoryless noisy channels. In *IEEE International Conference on Communications*, 1988.

[154] J. Zheng and B. D. Rao. Analysis of MIMO with finite-rate channel state information feedback: a source coding perspective. *IEEE Trans. on Signal Processing*, Vol. 55:4612–4626, Sep. 2007.