

CAPACITY ANALYSIS OF
AMPLIFY-AND-FORWARD MULTI-ANTENNA
RELAY CHANNELS

CORSO DI LAUREA SPECIALISTICA IN
INGEGNERIA DELLE TELECOMUNICAZIONI

DIPARTIMENTO DI ELETTRONICA E INFORMAZIONE
POLITECNICO DI MILANO
A.A. 2005 - 2006

TESI DI:
Nicola Varanese
MAT. 667002

RELATORE: Chiar.mo Prof. Umberto Spagnolini
CO-RELATORE: Dott. Osvaldo Simeone
CO-RELATORE: Chiar.mo Prof. Yeheskel Bar-Ness

*This work is dedicated to my mother,
for her courage.*

Acknowledgements

I like to think of this thesis as the outcome of the events that crossed my life in the last six years or so. A large part of this events was nice and enjoyable. A few, it's better not to say.

I spent the first five years studying at the Politecnico di Milano, building the basis for my future, under the guidance of most valuable professors. Among the many, I have to first thank Prof. Umberto Spagnolini, for his wise advisorship in this and other works.

The last six months that I've spent at the Center for Wireless Communications and Signal Processing Research at the NJIT, have surely been great. Thus, I would like to thank Prof. Yeheskel Bar-Ness for his supervision and hospitality, and everyone in the CWCSRP crew, especially Cao Mingcheng for the interesting discussions.

Nothing of what happened to me here in the US would have been possible without the guidance and experience of Doc. Osvaldo Simeone: you really taught me something, I must say!

Among my many friends, Alessandro Barenghi is the first to ACK(nnowledge) for his most precious help and support during all my stay in the US. Many thanks to Lorenzo T. Armano, for his friendship and refined political talk; to Gabriele "the wizard" Palletta for his knowledge; to Riccardo Repetto for his loyalty and to Amedeo Carnevali for his bravery.

Finally, I have to thank Alessandra for her loving support during all these years, and my mother, to whom this work is dedicated.

Contents

Acknowledgements	v
1 Introduction	1
1.1 The MIMO channel	2
1.2 Multi-user and cooperative communication	5
1.3 The Relay Channel	6
1.3.1 Introduction of Practical Constraints	8
1.4 Main contributions	11
2 The Multi-hop MIMO AF Relay Channel	15
2.1 System model	15
2.2 Problem formulation	16
2.3 An iterative solution	17
2.3.1 The optimization of the source input covariance matrix	17
2.3.2 The optimization of the linear processing at the relay	18
2.3.3 Algorithm definition	19
2.3.4 Simulation results	20
2.4 Low-complexity implementation	21
2.4.1 Analysis of the solution for the relay processing	22
2.4.2 Constant-power water-filling	26
2.4.3 A duality gap bound on the accuracy of constant-power water-filling	28
2.4.4 Low-complexity iterative solution	30

2.4.5	Simulation results	31
3	The AF Relay Channel with Multiple Antennas at the Relay	39
3.1	System model and problem definition	40
3.2	Optimization of the relay linear processing	44
3.3	Optimization of the power allocation	45
3.4	Simulations results	46
4	The Cooperative MIMO AF Relay Channel	51
4.1	System model	52
4.2	Problem formulation	54
4.3	An iterative solution	55
4.3.1	The optimization of the source input covariance matrix	56
4.3.2	The optimization of the linear processing at the relay	56
4.3.3	Algorithm definition	58
4.3.4	Simulation results	58
4.4	Low-complexity iterative solution	61
5	Conclusions	67
A	Convex Optimization Fundamentals	69
A.1	General optimization problems and the duality gap	69
A.2	Convex functions	70
A.3	Convex optimization problems	71
	Bibliography	73

Chapter 1

Introduction

In the past twenty years, the use of multiple antennas in radio transceivers has been regarded as the most promising technology for improving throughput, reliability and resistance to interference in wireless networks. Fig. 1.1 shows the downlink of a cellular system with an antenna array as an example. The deployment of multiple antennas enables the use of powerful signal processing techniques (from which the name *smart*-antennas) in order to improve the Signal-to-Interference-and-Noise-Ratio and thus the achievable rate (*array gain*). Furthermore, the availability of multiple copies of the received signal affected by independent fading processes leads to a more reliable transmission in terms of Bit-Error-Rate (*diversity gain*).

In recent years, the proposal of using multiple-antennas at the receiver side as well, has opened up a wealth of opportunities for performance improvement, as anticipated by information theoretic analysis. Such systems are referred to as Multiple-Input-Multiple-Output (MIMO). While the capacity of the MIMO channel has been derived in closed form for a variety of cases, the capacity of multi-user MIMO channels deserves further investigation.

The broadcast nature of the wireless channel, namely the possibility for each node to overhear transmissions from other nodes, enable the deployment of a variety of multi-user communication schemes. Of particular interest are the so called *collaborative* communications techniques, whereby different nodes of a wireless network cooperate, instead of competing, for radio resources.

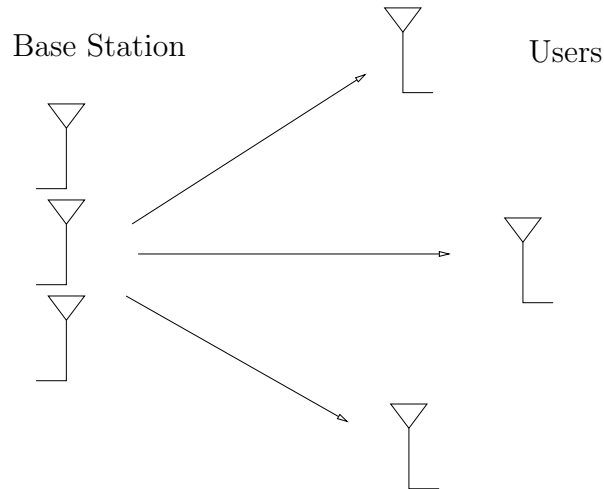


Figure 1.1: Cellular system with multiple antennas deployed at the base station

The topic of this thesis is the application of the multiple antenna technology to a cooperative scenario. In Sec. 1.1 key results for single-user MIMO communication systems are presented. In Sec. 1.2 an overview on classic results on multi-user and cooperative communications is presented, focusing on the relay channel in Sec. 1.3. Finally, in Sec. 1.4 we present the original contributions of this work.

1.1 The MIMO channel

In the early 90's, the capacity of a single-user Multiple-Input-Multiple-Output (MIMO) communication channel was derived in [1] [2]. A single-user MIMO system consists of a source and a destination, both equipped with multiple antennas. MIMO systems enjoy the typical benefits of multiple-antenna systems, namely the array-gain and the diversity gain. Moreover, the theoretical results promise huge performance boosts in terms of capacity, due to the possibility to decompose the channel into N parallel AWGN Single-Input-Single-Output (SISO) links exploiting the Singular Value Decomposition (SVD) of the channel matrix $\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^H$.

The capacity and optimal power allocation for a system where the source node could use at once N independent parallel channels to communicate with the destination is known since the early days of information theory [3]. The optimal power

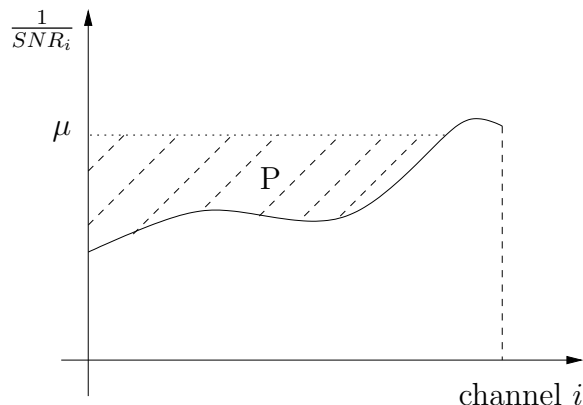


Figure 1.2: The water-filling power allocation.

allocation for such a system is the well-known *water-filling* scheme, illustrated fig. 1.2. Given the SNR_i for each channel, the power¹ p_i is allocated rising the “water-level” μ over SNR_i^{-1} until the total allocated power $\sum_i p_i = \sum_i \mu - SNR_i^{-1}$ is equal to the power constraint P . Using this power allocation, the capacity is just the sum of the rates over each channel $\sum_i \log(1 + p_i SNR_i)$. The model of parallel channels applies to communication systems that operate in both *time* and *frequency* domain.

In the time domain, it was shown in [4] that a SISO link affected by ergodic fading can be decomposed into N parallel channels, where N is the number of the channel states. In this case, the optimal power allocation requires water-filling in the time domain. It was also shown that the capacity of the equivalent parallel channels is greater than the capacity of an AWGN channel with the same average SNR . Unfortunately, the coding theorem used to demonstrate the achievability of the capacity requires the use of rather complicated transmitter structures, capable of changing the transmitted power and rate adaptatively according to the channel realization. Of course, in order to do that, the transmitter has to have full Channel State Information (CSI) fed back from the destination through a reliable, low-delay link [7].

On the other hand, in the frequency domain, communication systems based on parallel channels have been widely investigated since the 80’s, in order to combat

¹The powers are considered normalized upon 1 W, and thus are adimensional.

heavily frequency-selective channels. The principle is simply to use signal processing techniques (namely the Discrete Fourier Transform) to divide the frequency spectrum into N independent narrow sub-bands, on each of which the channel is constant. The optimal power allocation in this case is a water-filling in the frequency domain. This communication system is called in the literature Orthogonal Frequency Division Multiplexing (OFDM), and has been so far widely implemented in different applications, such as Digital Subscriber Lines (DSL) systems for broadband Internet access.

In the case of MIMO systems, the parallel channels are established in the *spatial* domain. This is a most welcomed feature, since in order to set up a new channel it is required only to deploy one more antenna at the transmitter *and* at the receiver, and no adaptative techniques or further bandwidth consumption are needed. Yet, the transmitter still has to have full CSI in order to choose the optimal signaling subspace and perform the optimal power allocation, which is a water-filling in the space domain. It has been shown in [1] that, if the SNR tends to infinity, the capacity of a MIMO channel scales linearly with the number of antennas N . This potentially huge capacity gain is called in the literature *multiplexing gain*.

A plethora of transmitter and receiver architectures have been recently proposed to attain the gains (multiplexing and diversity) promised by the theory. In [8] the concept of Space-Time-Block-Codes (STBC) was introduced to achieve the maximum degree of diversity of a MIMO channel. In particular, a code capable of achieving the full diversity gain of a MIMO channel with 2 transmitting and receiving antennas was presented. STBCs are a very simple transmission technique, since the optimal receiver structure is linear and no CSI at the transmitter is assumed. Several works followed, presenting different types of optimal and sub-optimal codes for different number of antennas. In [2], a transmitter and receiver architecture was proposed, capable of achieving the full multiplexing gain. On the transmitter side a linear transformation is needed to transmit the N independent coded streams on the correct subspace, while at the receiver side different structures can be implemented. The capacity achieving technique is a combination of Minimum Mean Square Error (MMSE) detection and Successive Interference Cancellation (SIC). Trade-off between multiplexing and diversity for MIMO systems has been thoroughly investigated in [9].

Finally, it should be emphasized that not only a high SNR is needed in practice to fully benefit of the multiplexing gain. In fact, if the channel matrix is rank-deficient, the diagonalized channel will have less than N channels. This situation is common in Line-of-Sight (LOS) situations, that is, when the propagation environment does not provide sufficiently rich scattering.

1.2 Multi-user and cooperative communication

Multi-user communication was introduced at the dawn of information theory by Shannon in his early studies [10] on the multiple access channel. The simple point-to-point (or one-sender-one-receiver) channel is encompassed by more complicated communication models where multiple senders are allowed to simultaneously use the channel resources to communicate with multiple receivers. In this setting, the set of all achievable rates from each sender to each receiver form the so-called *capacity region*. A lot of efforts were put in this field in the 70's and early 80's [14]. Unfortunately, the capacity region of a few multi-user channels is known, namely the Multiple Access Channel (MAC) [3] and the Broadcast Channel (BC) [11]. Different flavors of achievable rates have been found for the interference channel, [12], [13]. For more general communication networks, only a loose max-flow-min-cut upper bound has been derived so far [3].

Recently, this field of research has re-gained attention from the research community. In fact, the very nature of the wireless medium enables more complicated communication schemes than simple point-to-point or broadcast/multi-access channels. Each node in a wireless network can overhear the transmission of other nodes and, instead of rejecting this information as interference, the node could employ some more complicated source/channel coding scheme to improve the overall system performance.

From an information-theoretic point-of-view, the concept of cooperation is rather simple: each node decodes the message transmitted by the other nodes and retransmits it, in whole or in part, according to a network coding scheme. The simplest cooperative channel, the *relay channel* was first studied by Cover in [15] (see fig.

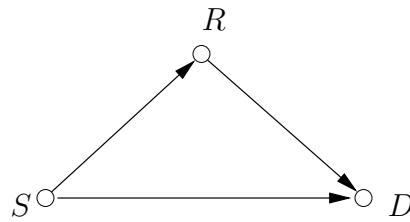


Figure 1.3: Relay System: source node (S), relay node (R) and destination node (D)

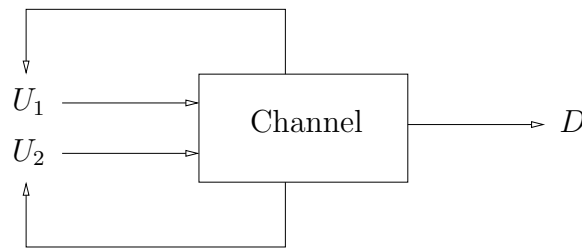


Figure 1.4: MAC with generalized feedback: user 1 (U_1), user 2 (U_2), and destination (D)

1.3). The relay channel is made up of three nodes: a source node, a relay node and a destination node. The information to be transmitted is originated at the source node and the relay node simply aids the communication between the source and the destination. Another early example of analysis of a cooperative system is found in the work of Carleial [16] and Willems [17], where it was shown that the capacity of a MAC (the simplest channel affected by interference) is increased not only by feedback from the receiver, but even by feedback from the channel itself (see fig. 1.4). A feedback link from the channel can be thought of as an observation of the interference present in a wireless environment, for example. In that situation a *generalized* feedback link comes totally for free, so why not taking advantage of it?

1.3 The Relay Channel

As discussed above, the simplest cooperative system is the relay channel in fig.1.5, first introduced by van der Meulen [18], and then thoroughly explored by Cover in his landmark paper [15]. This channel comes from the interaction between three terminals: a source node, from which information originates, a destination node,

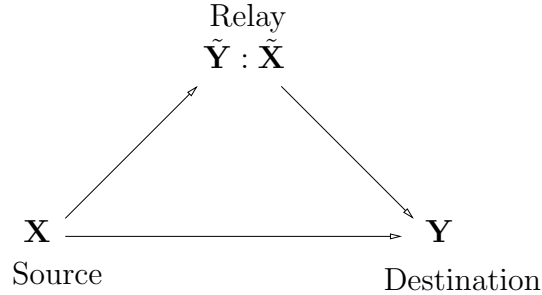


Figure 1.5: The relay channel as introduced by Van der Meulen.

to which information is destined, and a relay node, whose operation should yield substantial improvement to the overall system performance (in terms of throughput, reliability). Anyway, the capacity of this channel is known only for the physically degraded case, and only upper and lower (achievable rates) bounds have been found for the general setting. In this section we will present this channel starting from the most general vector model.

In the discrete-memoryless vector case, the relay channel is made up of four finite sets \mathcal{X} , $\tilde{\mathcal{X}}$, \mathcal{Y} and $\tilde{\mathcal{Y}}$, and a collection of probability mass functions (channel transition function) $p(\cdot, \cdot | \mathbf{x}, \tilde{\mathbf{x}})$ on $\mathcal{Y} \times \tilde{\mathcal{Y}}$, one for each $(\mathbf{x}, \tilde{\mathbf{x}}) \in \mathcal{X} \times \tilde{\mathcal{X}}$. The usual interpretation for the random variables in fig.1.5 is that \mathbf{X} is the input symbol to the channel as chosen by the source; \mathbf{Y} is the output of the channel as seen by the destination decoder; $\tilde{\mathbf{Y}}$ is the relay's observation, and finally $\tilde{\mathbf{X}}$ is the input symbol as chosen by the relay. In this Section, uppercase letters represent a random variable, while lower case letters represent the corresponding realization.

Here we give the most general possible definition for a (M, n) code for the relay channel, where n is the number of channel uses (code block length), $M = 2^{nR}$ is the number of source messages, and R is the code rate. Such a code would consist of a set of M source messages $\mathcal{W} = \{1, 2, \dots, M\}$; a channel encoding function (available at the source node)

$$\mathbf{X}^n : \mathcal{W} \rightarrow \mathcal{X}^n$$

which maps the message index w into a codeword $\mathbf{X}^n(w)$ composed by n vector symbols chosen from the input alphabet \mathcal{X} ; a set of relay (vector) functions $\{\mathbf{f}_i\}_{i=1}^n$

such that

$$\tilde{\mathbf{x}}_i = \mathbf{f}_i(\tilde{\mathbf{Y}}_1, \tilde{\mathbf{Y}}_2, \dots, \tilde{\mathbf{Y}}_{i-1}) \quad 1 \leq i \leq n \quad (1.1)$$

where i is the channel use index, and $\tilde{\mathbf{x}}_i \in \tilde{\mathcal{X}}$; and a decoding function (available at the destination)

$$g : \mathcal{Y}^n \rightarrow \mathcal{W}$$

which maps the received vector word into the corresponding message. The definition of the relay functions (1.1) stems from the usual *non-anticipatory* constraint over the relay operation, that is the fact that the relay output at time instant i should depend *causally* over the relay output symbols $\{\tilde{\mathbf{Y}}_i\}_{i=1}^{i-1}$. Given this condition, the channel is memoryless in the sense that $(\mathbf{Y}_i, \tilde{\mathbf{Y}}_i)$ is independent from $\{(\mathbf{X}_j, \tilde{\mathbf{X}}_j)\}_{j \neq i}$ (i.e. past and future symbols) given $(\mathbf{X}_i, \tilde{\mathbf{X}}_i)$. Thus the channel transition function for n uses of the channel can be expressed as

$$p(\mathbf{y}^n, \tilde{\mathbf{y}}^n | \mathbf{x}^n, \tilde{\mathbf{x}}^n) = \prod_{i=1}^n p(\mathbf{y}_i, \tilde{\mathbf{y}}_i | \mathbf{x}_i, \tilde{\mathbf{x}}_i)$$

and for *any* choice of $p(w)$, $w \in \mathcal{W}$, channel code $\mathbf{X}^n : \mathcal{W} \rightarrow \mathcal{X}^n$, and relay functions $\{\mathbf{f}_i\}_{i=1}^n$, the joint probability mass function on $\mathcal{W} \times \mathcal{X}^n \times \tilde{\mathcal{X}}^n \times \mathcal{Y}^n \times \tilde{\mathcal{Y}}^n$ is given by:

$$p(w, \mathbf{x}^n, \tilde{\mathbf{x}}^n, \mathbf{y}^n, \tilde{\mathbf{y}}^n) = p(w) \prod_{i=1}^n p(\mathbf{x}_i | w) p(\tilde{\mathbf{x}}_i | \tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2, \dots, \tilde{\mathbf{y}}_{i-1}) p(\mathbf{y}_i, \tilde{\mathbf{y}}_i | \mathbf{x}_i, \tilde{\mathbf{x}}_i)$$

1.3.1 Introduction of Practical Constraints

Cover [15] found a max-flow-min-cut upper bound for the capacity of the relay channel as presented above. He also demonstrated that, for the case of degraded relay channels, this upper bound boils down to the rate achievable with the Block-Markov Coding (BMC) scheme, also presented in his paper. Anyway this coding technique is currently regarded as too complicated from an implementation standpoint. Furthermore, in the above theoretical framework, most of the constraints of real-world systems are missing.

In fact, in [15] the relay is assumed to be able to receive and transmit at the same

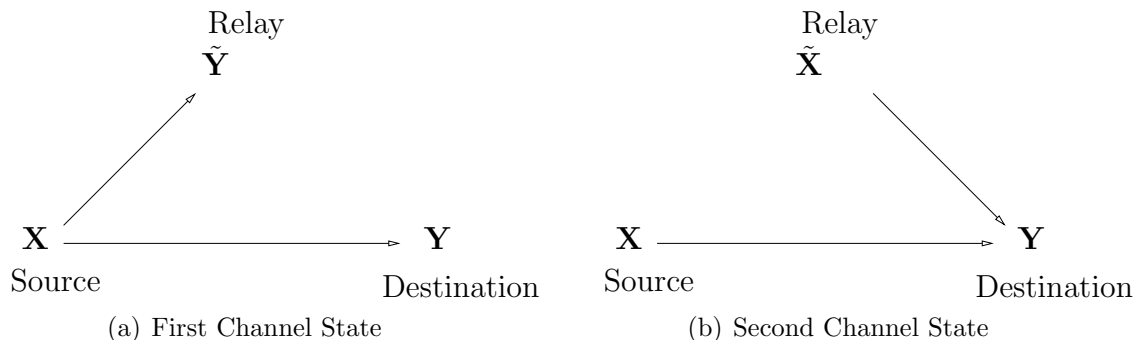


Figure 1.6: Multistate relay channel

time and in the same frequency band (Full Duplex relay operation), which is quite a difficult condition to realize in practice, due to the enormous difference in the power levels of the received and transmitted signals at the relay station. So practicality imposes the constraint that the relay cannot receive and transmit signals on the same frequency band at the same time instant: the relay will have to use two different orthogonal channels for transmission and reception (*Half Duplex* constraint). From an information-theoretical point-of-view this constraint changes the nature of the channel. This new channel can be modeled in two different ways. In fact, it can be regarded as a *two-state* channel: in the first state (fig. 1.6(a)) the relay listens to the source, which is transmitting information to both the relay and the destination node; in the second state (fig. 1.6(b)) the relay transmits to the destination a codeword chosen according to its relaying functions (of course the *non-anticipatory* condition still holds) while the source node transmits a new message. This model is also called in the literature “*cheap*” relay channel and has been investigated in [19] [20] applying the usual max-flow-min-cut theorem introduced by Cover [3]. Another possible model is the *parallel* relay channel [21], but the results are identical. We choose to orthogonalize the relay operation in the time domain (TDD - Time Division Duplex relay operation), regarding the two possible channel states respectively as the transmission scheme in the first and in the second time-slot. Generally speaking, the relative duration of the two time-slots could be different and subject to optimization, but, depending on the nature of the relaying functions, it could also be fixed.

The relaying function can be chosen according to system design requirements: it

could be linear or non-linear, a decoding or a signal-processing function, it could be even a stochastic function [15]. In this work, the most practical solution is investigated: the relay applies a linear transformation to the signal received in the first time slot (i.e., a simple scaling in the scalar case). Notice that this choice also constrains the two time slots to have exactly the same length in signaling intervals. Thus, for the sake of simplicity, we assume that the duration of each time-slot will be just one signaling interval (i.e., and without loss of generality, the nodes transmit one symbol per time-slot). We can then rewrite (1.1) as

$$\tilde{\mathbf{x}}_2 = \mathbf{f}_2(\tilde{\mathbf{Y}}_1) = \mathbf{G}\tilde{\mathbf{y}}_1, \quad (1.2)$$

where the subscripts refer to the time-slot index. Notice that with this choices we do not put any sort of decoding/coding intelligence in the relay node, taking a step back from the complicated network coding schemes presented in [15]. A similar scheme was introduced by Laneman in [22] for the fading single-antenna case, and it is commonly referred to as the *Amplify-and-Forward* relaying protocol. Actually, Laneman’s protocol does not adhere strictly to the “cheap” relay model in that the source node is not allowed to transmit during the second time-slot. In fact, the second time-slot is totally dedicated to the relay transmission, and the source remains silent. Laneman adopted this simple scheme because he was mostly worried about diversity, and his AF protocol is actually able to achieve full asymptotic diversity gain. Instead, in this work the achievable rate (throughput) is considered as an important performance metric. Therefore, in order not to waste system resources, the source is allowed to transmit during the second time-slot a symbol \mathbf{x}_2 , which we suppose to be independent from the first time-slot transmitted symbol \mathbf{x}_1 . This is a slightly modified version of the original AF protocol first presented in [6], and indeed resembles the “cheap” relay model. In [6] it was also demonstrated that, without any joint relay/source power constraint, the throughput of this protocol is indeed higher than the protocol in [22]. The achievable rate of the AF protocol is easily shown to be

$$R_{AF} = \frac{1}{2}I(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y}_1, \mathbf{y}_2) \quad (1.3)$$

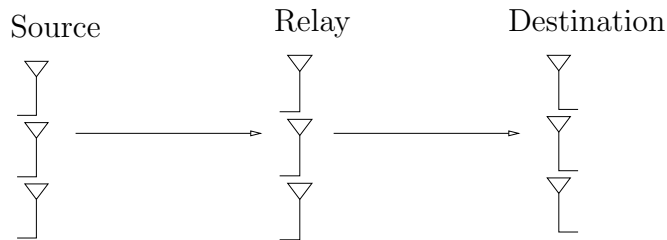


Figure 1.7: Multi-hop MIMO relay channel.

where $\frac{1}{2}$ accounts for the TDD operation. Notice that the rate (1.3) assumes that the destination node jointly decodes the first time-slot source symbol \mathbf{x}_1 and the second time-slot source symbol \mathbf{x}_2 , combining the signals received during *both* of the time-slots, \mathbf{y}_1 , \mathbf{y}_2 , and that the information about the linear transformation used at the relay is known to the source itself (full channel state information, CSI).

So far, MIMO cooperative communication systems have been investigated only in a few papers. The authors of [30] extended the information-theoretic results of [15] to a multi-antenna scenario and devised an algorithm to compute the input covariance matrices that maximize the max-flow-min-cut upper bound. Finally, performance of the AF scheme of [22] in a multi-antenna setting was analyzed in [5].

1.4 Main contributions

In this work, the potential benefits of deploying multiple antennas at the nodes of an AF relay channel are investigated.

In Chapter 2 the *multi-hop MIMO AF relay channel* is presented. An illustration of the system is shown in fig. 1.7. This system is *not* a cooperative communication system, strictly speaking. In fact, differently from [22] and [6], the destination remains idle during the first time-slot. In the first time-slot the source transmits to the relay, which retransmits the received symbol after a linear transformation to the destination in the second time-slot. Thus, there is no direct link between the source and the destination. For the multi-hop case, the problem of maximizing the achievable rate over the covariance matrix of the symbol transmitted by the source and relay linear processing matrix is formulated under the assumption of full channel state information

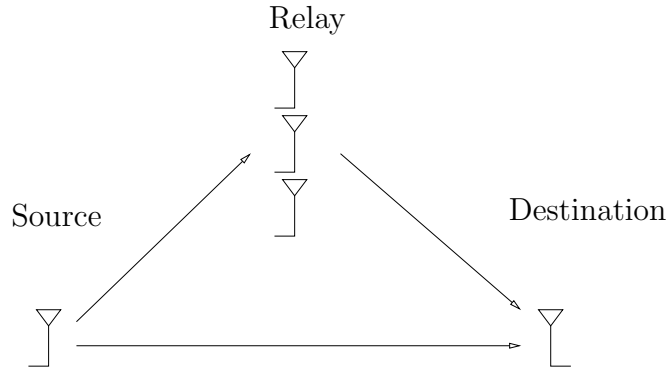


Figure 1.8: Cooperative relay channel with multiple antennas deployed at the relay node.

at each node. A sub-optimal iterative algorithm is proposed and proved by numerical simulations to outperform known schemes under particular channel circumstances. In particular, in each iteration of the algorithm, the source covariance matrix is optimized keeping the relay processing matrix fixed and vice-versa. A more accurate analysis of the relay optimization step enables a low-complexity implementation of the overall iterative algorithm. Also, the performance after just one-iteration of the low-complexity algorithm is shown to be not far from the rate achieved by the exact iterative algorithm.

In Chapter 3 we consider a fading *AF relay channel*, where each node has full CSI and *multiple-antennas are deployed at the relay node* as in fig 1.8. Maximization of the achievable rate with respect to the linear processing at the relay and the power allocation at the source and relay is performed under an instantaneous sum-power constraint. In particular, it is assumed that the total power used by all the active nodes during each time slot is fixed. It is proved that under such constraints, the scheme in [6] is never optimal, and the optimal strategy consists of using either direct transmission or the scheme in [22], according to which of the two achieves the higher rate.

In Chapter 4, the *cooperative MIMO AF relay channel* is presented. The system is sketched in fig. 1.9. As for the multi-hop case in Chapter 2, the problem of maximizing the achievable rate over the covariance matrices of the symbols transmitted by the source and relay linear processing matrix is formulated under the assumption of full

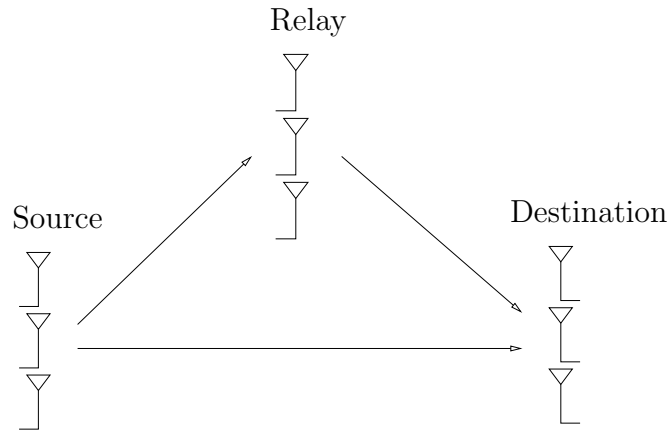


Figure 1.9: Cooperative MIMO relay channel.

channel state information at each node. Again, a sub-optimal iterative algorithm of the same fashion is proposed and proved by numerical simulations to outperform known schemes. In particular, in each iteration of the algorithm, the source covariance matrices are optimized keeping the relay processing matrix fixed and vice-versa. A low-complexity implementation of the algorithm is also proposed. The performance after just one-iteration of the low-complexity algorithm is shown to be not far from the rate achieved by the exact iterative algorithm. Finally, we draw our conclusions in Chapter 5. In Appendix A we present classic results and tools from convex optimization that will be used in the following discussion.

Chapter 2

The Multi-hop MIMO AF Relay Channel

The first system we analyze is the multi-hop MIMO AF relay channel. Strictly speaking, this system is not cooperative in the sense usually employed in the literature, since there is no direct exchange of messages between the source and the destination. However, as discussed in the following, many of the technical challenges of a cooperative system arise in this scenario as well.

In Sec. 2.1 the system model is presented. The optimization problem we tackle is detailed in Sec. 2.2, whereas an iterative sub-optimal solution is proposed in Sec. 2.3. Finally, in Sec. 2.4, a low-complexity iterative algorithm is devised in order to approximate the exact solution of the iterative algorithm defined in Sec. 2.3. Performance of this low-complexity solution is investigated through theoretical bounds (duality gap, see Sec. 2.4.3) and numerical results.

2.1 System model

The multi-hop MIMO Amplify-and-Forward relay is illustrated in fig.2.1. In this system three nodes are involved in the communication: a source, a relay and destination, each equipped with N antennas. The operation is divided into two time-slots: in the first time-slot the relay node receives the vector symbol transmitted by the source,

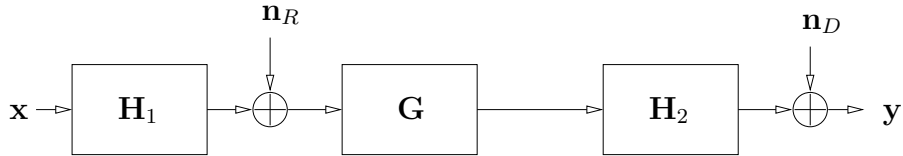


Figure 2.1: The multi-hop AF relay system.

while in the second time-slot the relay re-transmits the received vector symbol towards the destination after a linear transformation. The destination decodes based only on the signal received in the second time-slot. All the nodes are assumed to have full CSI. Moreover, the channel matrices are assumed to be independent Rayleigh fading processes, with a coherence time of at least two time-slots and i.i.d. entries.

During the first time slot, the relay receives a signal $\mathbf{y}_R = \mathbf{H}_1\mathbf{x} + \mathbf{n}_R$, where \mathbf{x} is the $N \times 1$ vector transmitted by the source, \mathbf{H}_1 is the $N \times N$ source-relay channel matrix and \mathbf{n}_R is the $N \times 1$ noise vector at the relay node, assumed to have distribution $\mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$. The relay processes \mathbf{y}_R through multiplication by a $N \times N$ matrix \mathbf{G} , retransmitting the symbol $\mathbf{x}_R = \mathbf{G}\mathbf{y}_R$ in the second time-slot. Therefore, the input/output relation for the overall system is:

$$\begin{aligned} \mathbf{y} &= \mathbf{H}_2\mathbf{x}_R + \mathbf{n}_D = \\ &= \mathbf{H}_2\mathbf{G}\mathbf{H}_1\mathbf{x} + \mathbf{H}_2\mathbf{G}\mathbf{n}_R + \mathbf{n}_D, \end{aligned} \quad (2.1)$$

where \mathbf{y} is the received symbol at the destination, \mathbf{H}_2 is the $N \times N$ relay-destination channel matrix, and \mathbf{n}_D is the $N \times 1$ noise vector at the destination, assumed to have distribution $\mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$.

2.2 Problem formulation

In this paper, we are interested in maximizing the mutual information between the source input \mathbf{x} and the destination output \mathbf{y} , $I(\mathbf{x}; \mathbf{y})$, over the source input covariance matrix $\mathbf{Q} = E[\mathbf{x}\mathbf{x}^H]$ and the signal processing at the relay \mathbf{G} , under instantaneous power constraints over the source and relay input symbols. Notice that the ergodic

achievable rate [1] of the system is $\frac{1}{2}E_{\mathcal{H}}[I(\mathbf{x}; \mathbf{y})]$, where $E_{\mathcal{H}}[\cdot]$ denotes the average with respect to fading and the factor $1/2$ accounts for the two-slots transmission. Thus, we can formulate our optimization problem as

$$\max_{\mathbf{Q}, \mathbf{G}} I(\mathbf{x}; \mathbf{y}) \quad (2.2a)$$

$$\text{s.t.} \begin{cases} \mathbf{Q} \succeq \mathbf{0} \\ \text{tr}(\mathbf{Q}) = P \\ \text{tr}(\mathbf{R}(\mathbf{Q}, \mathbf{G})) = P_R \end{cases} \quad (2.2b)$$

where

$$I(\mathbf{x}, \mathbf{y}) = C(\mathbf{H}_2 \mathbf{G} \mathbf{H}_1 \mathbf{Q} \mathbf{H}_1^H \mathbf{G}^H \mathbf{H}_2^H (\sigma^2 \mathbf{I} + \sigma^2 \mathbf{H}_2 \mathbf{G} \mathbf{G}^H \mathbf{H}_2^H)^{-1}), \quad (2.3)$$

$$\mathbf{R}(\mathbf{Q}, \mathbf{G}) = E[\mathbf{x}_R \mathbf{x}_R^H] = \mathbf{G}(\mathbf{H}_1 \mathbf{Q} \mathbf{H}_1^H + \sigma^2 \mathbf{I}) \mathbf{G}^H. \quad (2.4)$$

To simplify the notation, we define $C(\mathbf{X}) := \log(\det(\mathbf{I} + \mathbf{X}))$. According to (2.2b), the power is fixed to P for the source and to P_R for the relay node.

2.3 An iterative solution

The optimization problem (2.2) is not convex. However, if we fix either of the two matrix variables, \mathbf{Q} or \mathbf{G} , the resulting problem is convex in the remaining variable. Our solution to the problem (2.2) is then an iterative procedure that alternates between the optimization over \mathbf{G} fixed \mathbf{Q} and the optimization over \mathbf{Q} fixed \mathbf{G} . Absolute convergence to an optimal solution cannot be proved for this algorithm, since the power constraint on the relay input symbol depends upon *both* \mathbf{Q} and \mathbf{G} [31]. Nevertheless, if the problem is well-conditioned the algorithm has shown in practice rapid convergence to a unique sub-optimal solution.

2.3.1 The optimization of the source input covariance matrix

Let us start by fixing \mathbf{G} . The resulting optimization problem over \mathbf{Q} has a concave objective function and two affine equality constraints. However, as detailed in the following, for the sake of our algorithm, the second constraint can be ignored. It

follows that the resulting problem is

$$\max_{\mathbf{Q}} C(\mathbf{P}\mathbf{Q}\mathbf{P}^H) \quad (2.5a)$$

$$\text{s.t.} \begin{cases} \mathbf{Q} \succeq \mathbf{0} \\ \text{tr}(\mathbf{Q}) = P \end{cases}, \quad (2.5b)$$

where $\mathbf{P} = (\sigma^2\mathbf{I} + \sigma^2\mathbf{H}_2\mathbf{G}\mathbf{G}^H\mathbf{H}_2^H)^{-\frac{H}{2}}\mathbf{H}_2\mathbf{G}\mathbf{H}_1$ (and the matrix square-root is defined as $\mathbf{F} = \mathbf{F}^{\frac{H}{2}}\mathbf{F}^{\frac{1}{2}}$). Solution of (2.5) can be found according to [1] by transmitting along the eigenmodes of the equivalent channel \mathbf{P} , with power distributed along the sub-channels following the water-filling procedure.

2.3.2 The optimization of the linear processing at the relay

On the other end, if we fix \mathbf{Q} , the optimization problem boils down to

$$\begin{aligned} \max_{\mathbf{G}} \{ & C(\mathbf{H}_2\mathbf{G}\mathbf{A}\mathbf{A}^H\mathbf{G}^H\mathbf{H}_2^H(\sigma^2\mathbf{I} + \sigma^2\mathbf{H}_2\mathbf{G}\mathbf{G}^H\mathbf{H}_2^H)^{-1}) \} \\ \text{s.t.} \quad & \text{tr}(\mathbf{G}(\mathbf{A}\mathbf{A}^H + \sigma^2\mathbf{I})\mathbf{G}^H) = P_R, \end{aligned} \quad (2.6)$$

where $\mathbf{A} = \mathbf{H}_1\mathbf{Q}^{\frac{H}{2}}$ is the equivalent first-hop channel. This problem has been solved in [5]. Below we briefly summarize the solution by casting it into our notation.

Expanding \mathbf{H}_2 and \mathbf{A} with the corresponding singular value decomposition $\mathbf{H}_2 = \mathbf{U}_2\mathbf{\Lambda}_2\mathbf{V}_2^H$ and $\mathbf{A} = \mathbf{U}_A\mathbf{\Lambda}_A\mathbf{V}_A^H$, the objective and constraint functions are easily diagonalized by choosing $\mathbf{G} = \mathbf{V}_2\mathbf{D}_g\mathbf{U}_A^H$, where \mathbf{D}_g is a $N \times N$ diagonal matrix. Thus we can write the diagonalized optimization problem in standard form (as defined in [28])

$$\begin{aligned} \min_{\{g_1 \dots g_N\}} \left\{ - \sum_{r=1}^N \log \left(1 + \frac{\lambda_{A,r}^2 \lambda_{2,r}^2 |g_r|^2}{\sigma + \sigma \lambda_{2,r}^2 |g_r|^2} \right) \right\} \\ \text{s.t.} \begin{cases} -|g_r|^2 \leq 0 & r = 1, \dots, N \\ \sum_{r=1}^N (\lambda_{A,r}^2 + \sigma) |g_r|^2 = P_R \end{cases}, \end{aligned} \quad (2.7)$$

where $\lambda_{A,r}$ and $\lambda_{2,r}$ are the r -th singular value of \mathbf{A} and \mathbf{H}_2 and g_r is the r -th diagonal

element of the matrix \mathbf{D}_g . This problem can be solved analytically using the Karush-Kuhn-Tucker conditions [28]. The solution is similar to a water-pouring over the eigenmodes of the relay-destination channel \mathbf{H}_2 :

$$|g_r|^2 = \frac{1}{(\lambda_{A,r}^2 + \sigma^2)} \left[f(\mu; \eta_r) - \frac{\sigma^2}{\lambda_{2,r}^2} \right]^+ \quad (2.8)$$

$$f(\mu; \eta_r) = \sqrt{\left(\frac{\eta_r}{2}\right)^2 + \eta_r \mu} - \frac{\eta_r}{2}, \quad (2.9)$$

where $[x]^+ = \max(x, 0)$ and $\eta_r = \frac{\lambda_{A,r}^2}{\lambda_{2,r}^2}$. The value of μ is chosen as to satisfy the power constraint on the relay input symbol $\sum_{r=1}^N (\lambda_{A,r}^2 + \sigma^2) |g_r|^2 = P_R$.

2.3.3 Algorithm definition

Following from this results, we can finally detail our algorithm:

```

initialize  $\mathbf{Q} = \mathbf{0}$  ,  $\mathbf{G} = \mathbf{I}$ ;
repeat
(a) solve the optimization problem (2.5) for  $\mathbf{Q}$  keeping  $\mathbf{G}$  fixed;
(b) solve the optimization problem (2.6) for  $\mathbf{G}$  keeping  $\mathbf{Q}$  fixed;
until the rate (2.3) converges.

```

We choose to initialize $\mathbf{G} = \mathbf{I}$, because otherwise there would be no connection between the source and the destination at the starting point of the optimization. Notice that the last iteration of the algorithm has to be (2.6), since this guarantees the enforcement of all the constraints in the original problem (2.2). As we said before, absolute convergence to an optimal solution is not guaranteed for this algorithm. However, if the problem is well-conditioned, it has shown to rapidly converge¹ to a unique sub-optimal solution even from a randomly chosen starting point. This algorithm is efficient in that it decomposes the general non-convex multi-user problem into a sequence of convex single-user problems, each of which is much simpler to solve.

¹around 5-6 iterations with a tolerance of 10^{-2} on the rate.

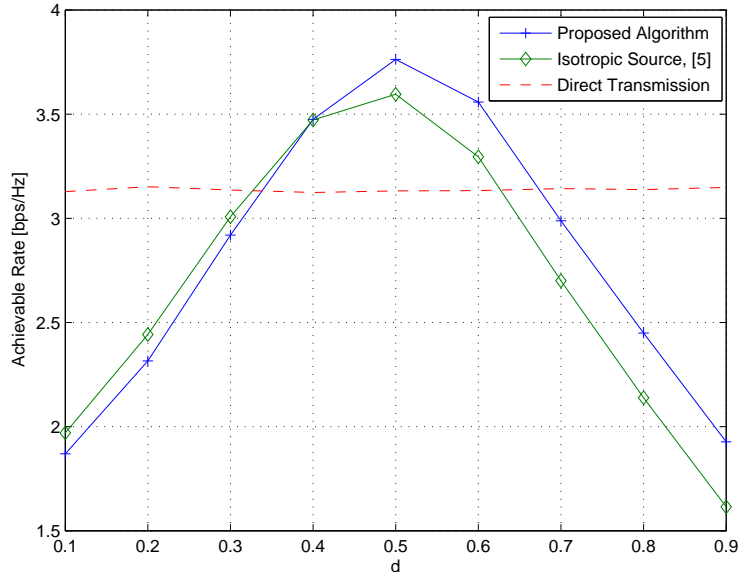


Figure 2.2: Achievable rates of different communication schemes for the multi-hop system with $\frac{1}{\sigma^2} = 0dB$, $N = 3$.

2.3.4 Simulation results

In this section, we assume that the relay is located on a line between the source and the destination, at a normalized distance $d \in [0, 1]$ from the source and $(1 - d)$ from the destination. It follows that, assuming a path-loss exponent of 4, the channel matrix \mathbf{H}_i has entries distributed as

$$[\mathbf{H}_i]_{m,n} \sim \mathcal{CN}(0, d_i^{-4}), \quad (2.10)$$

where $d_1 = d$, $d_2 = (1 - d)$ and $d_0 = 1$. As far as the power constraints are concerned we fixed both P and P_R to unity. The ergodic achievable rates of different algorithms for the multi-hop system are depicted in fig. 2.2. In particular, both the technique presented in [5] (that assumes an isotropic covariance matrix $\mathbf{Q} = N^{-1}\mathbf{I}$) and the proposed method are compared to the reference performance of direct transmission (between source and destination). Multi-hop transmission yields a performance gain (up to 0.5 bps/Hz) for a large range of values of d , and, in this range, the proposed

algorithm performs better than previously proposed solution. Further simulation results are presented in Sec. 2.4.

2.4 Low-complexity implementation

In Sec. 2.3 we devised an iterative procedure in order to find a suboptimal solution to the general optimization problem (2.2). In each step of the algorithm a single-user water-filling is performed to find the eigenvalues of the matrix \mathbf{Q} ; then, a generalization of the same water-filling algorithm is performed to find the singular values of the relay processing matrix \mathbf{G} . The power allocation relative to the optimization of \mathbf{G} can be quite expensive in terms of computational cost, since it requires to compute several times the value of the non linear function $f(\mu; \eta_r)$ in (2.8).

In this Section, we tackle the problem of finding a *low-complexity* algorithm to approximate the *exact* solution of the iterative procedure presented above. The first problem to solve in this direction is to simplify the optimization of the relay processing matrix \mathbf{G} . In particular, in Sec. 2.4.1 we analyze the function $f(\mu; \eta_r)$ and develop some intuition on the exact solution of (2.8). We will show that, assuming the geometrical model introduced in Sec. 2.3.4, the solution of (2.8) can be approximated with a direct water-filling allocation on the modes of the second-hop with a small performance loss. In order to further reduce the computational burden, a constant-power water-filling allocation over the modes of the second hop is proposed in Sec. 2.4.2. In Sec. 2.4.3, an analytic bound on the accuracy of constant-power allocation is derived using convex analysis. Finally, in Sec. 2.4.4, we detail the low-complexity iterative algorithm, using constant-power water-filling allocation to perform both the optimization of \mathbf{Q} and the optimization of \mathbf{G} . Further simulation results are presented in Sec. 2.4.5, comparing the achievable rates of the exact and low-complexity iterative algorithm. Also, the achievable rates after just one iteration of the low-complexity algorithm are compared to the performances at the convergence of the exact algorithm.

2.4.1 Analysis of the solution for the relay processing

In order to have more insight on the solution of the problem (2.7), it is useful to reformulate (2.8) as

$$p_r = (\lambda_{A,r}^2 + \sigma^2)|g_r|^2 = \left[f(\mu; \eta_r) - \frac{\sigma^2}{\lambda_{2,r}^2} \right]^+, \quad (2.11)$$

where p_r is the power allocated to the r -th sub-channel and the relay power constraint expressed in terms of p_r is simply $\sum_{r=1}^N p_r = P_R$. This expression is indeed similar to a conventional water-filling on the channel \mathbf{H}_2 . In fact, we can distinguish two terms in (2.11): the last term determines the “depth” of the r -th bin, while the first term is the “water level” upon the r -th bin. The last term is the same we would find if we chose $|\tilde{g}_r|^2$ with a water-filling procedure over the eigenmodes of \mathbf{H}_2 . Differently from conventional water-filling, the first term (the water level) is not equal for all the sub-channels. In fact it is a parametric function of μ (2.9).

The function $f(\mu; \eta_r)$ is plotted in fig. 2.3 for different values of the parameter η_r . We see that the water-level $f(\mu; \eta_r)$ is a nonnegative monotonically increasing function of μ , which means that increasing the Lagrangian multiplier μ we pour more water over the second-hop bins. Yet, the resulting water level is different for each bin, depending on η_r . For a fixed value of μ , the water level is higher on the bins which corresponds to higher values of the parameter η_r . This means that the relay tries to compensate for the possible loss in the rate supported by a specific sub-channel due to a poor second-hop power gain. On the contrary, if the second-hop power gain is much stronger than the first-hop power gain, the relay allocates power slower. If η_r was constant, then the water-level would be equal on all the sub-channels, and the optimal relay power allocation would be the water-filling on the eigenmodes of \mathbf{H}_2 . The Taylor-series expansion of the water-level function $f(\mu; \eta_r)$ is found to be

$$f(\mu; \eta_r) = \mu + \eta_r \sum_{i=2}^{+\infty} k_i \left(\frac{\mu}{\eta_r} \right)^i. \quad (2.12)$$

Thus, it is easy to see that for small values of μ $f(\mu; \eta_r) \approx \mu$, and the water-level is

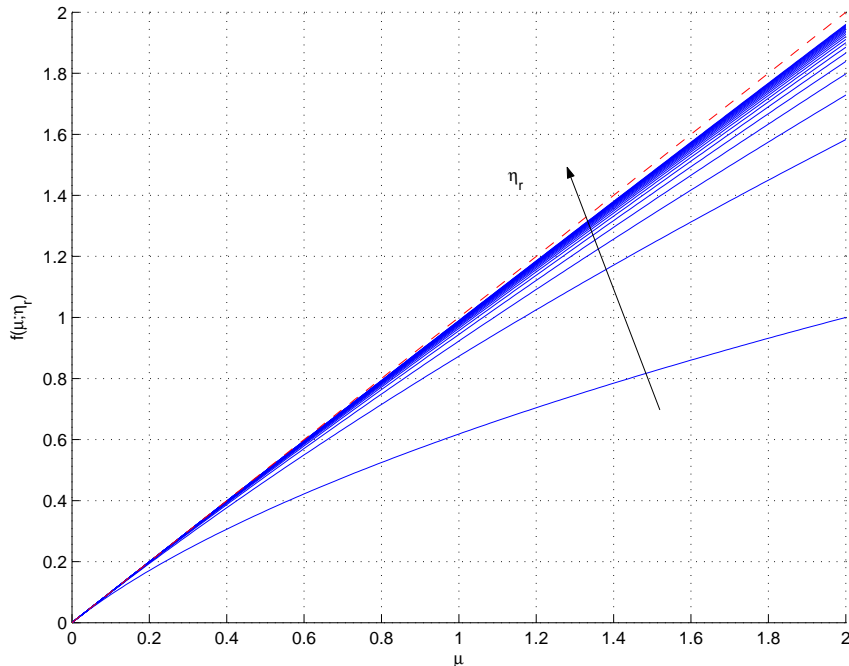


Figure 2.3: The function $f(\mu; \eta_r)$ for different values of η_r .

almost the same at the beginning of the water-filling procedure. If $\lambda_{2,r} = 0$, that is in the limit of $f(\mu; \eta_r)$ over $\eta_r \rightarrow \infty$, from (2.12) it is straightforward to prove

$$\lim_{\eta_r \rightarrow \infty} f(\mu; \eta_r) = \mu. \quad (2.13)$$

This result upper-bounds the water-level on sub-channels where the first-hop power-gain is much stronger than the second-hop power gain. From fig. 2.3 it seems also that the convergence to this limit is pretty fast. However, as expected, if the second-hop was really poor, the depth $\sigma^2/\lambda_{2,r}^2$ would go to infinite and no power would be allocated on that sub-channel by the relay. On the other hand, if $\lambda_{A,r} = 0$ for some r , $\eta_r = 0$ and the water-level would be zero for every value of μ and no power would be allocated to that sub-channel for transmission on the second hop.

Now it is useful to reformulate the channel model introduced in Sec. 2.3.4 as in the following

$$\mathbf{H}_i = \frac{1}{d_i^2} \mathbf{H}_{w,i}, \quad (2.14)$$

where each entry of $\mathbf{H}_{w,i}$ is i.i.d. with distribution $\mathcal{CN}(0, 1)$. The parameter η_r , defined as the ratio of the squared singular values of \mathbf{A} and \mathbf{H}_2 , can be expressed as a function of the model parameter d

$$\eta_r = \underbrace{\left(\frac{1-d}{d}\right)^4}_{k(d)} \underbrace{\frac{\delta_{A,r}}{\delta_{2,r}}}_{\tilde{\eta}_r}. \quad (2.15)$$

The first factor $k(d)$ in (2.15) is deterministic and depends on the distance between the relay and the destination, while the second term, $\tilde{\eta}_r$, is the ratio of two random variables, namely the r -th eigenvalues of the matrices $\mathbf{H}_{w,1}\mathbf{Q}\mathbf{H}_{w,1}^H$ and $\mathbf{H}_{w,2}\mathbf{H}_{w,2}^H$. Since under the hypothesis of Rayleigh i.i.d. fading processes $\mathbf{H}_{w,2}\mathbf{H}_{w,2}^H$ is distributed according to the Wishart distribution, the joint pdf of its eigenvalues is known to be

$$P_2^{n,k}(\delta_{2,1}, \dots, \delta_{2,k}) = \prod_i^k \delta_{2,i}^{n-k} e^{-\delta_{2,i}} \times \prod_{i<j}^k (\delta_{2,i} - \delta_{2,j})^2. \quad (2.16)$$

On the other hand, it seems hard to find the joint pdf of the eigenvalues of $\mathbf{H}_{w,1}\mathbf{Q}\mathbf{H}_{w,1}^H$, since \mathbf{Q} is chosen in each iteration as the water-filling over the equivalent channel $(\sigma^2\mathbf{I} + \sigma^2\mathbf{H}_2\mathbf{G}\mathbf{G}^H\mathbf{H}_2^H)^{-\frac{H}{2}}\mathbf{H}_2\mathbf{G}\mathbf{H}_1$. Thus, it seems all the more difficult to characterize the statistics of $\tilde{\eta}_r$. The deterministic factor $k(d)$, on the contrary, has a very simple expression, as we assumed a one-dimensional geometric model. This factor has been plotted for different values of the path-loss exponent in fig. (2.4). Since $k(d)$ goes rapidly to $+\infty$ when the relay is close to the source and to 0 when the relay is close to the destination, it turns out that $\tilde{\eta}_r$ has a minor impact in determining the value of η_r . Thus, we can approximate the water-level over each bin as

$$f(\mu; \eta_r) \approx f\left(\mu; \left(\frac{1-d}{d}\right)^4\right). \quad (2.17)$$

A graphic representation of this approximation is depicted in fig. 2.5. As discussed above, if the water-level is the same over each bin the algorithm boils down to the simple single-user water-filling.

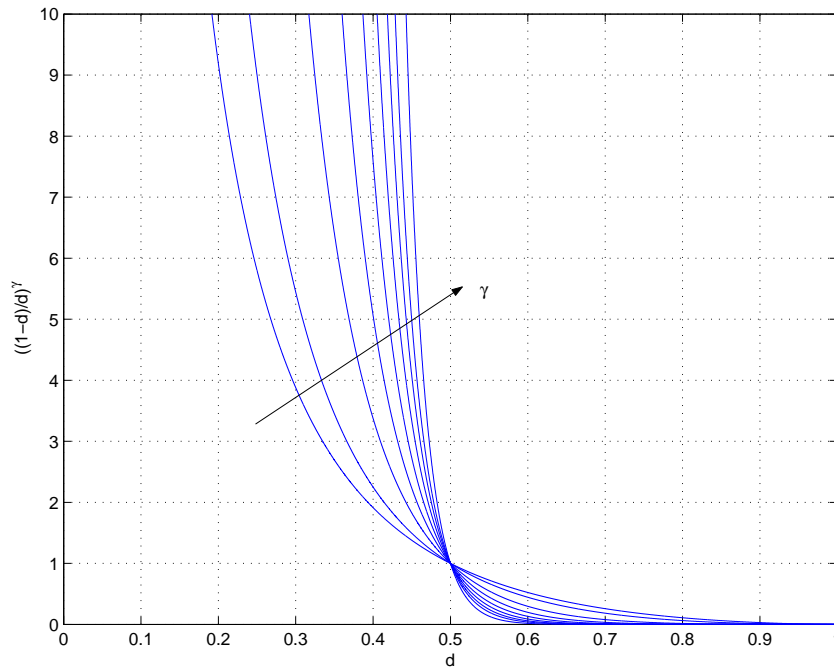


Figure 2.4: The geometric factor $((1-d)/d)^\gamma$ for different values of the path-loss exponent γ .

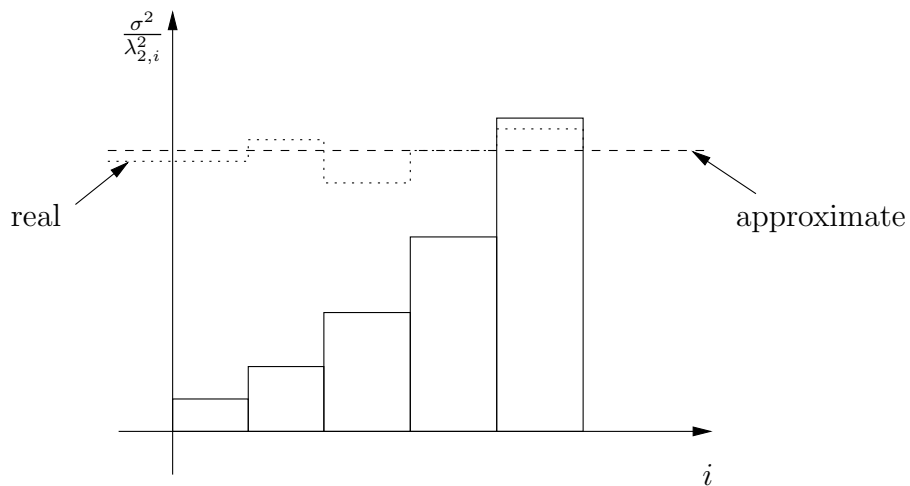


Figure 2.5: The approximate and real water-levels over $\frac{\sigma^2}{\lambda_{2,r}^2}$.

Building on this intuition, a simple sub-optimal algorithm to compute \mathbf{G} is to allocate the powers p_r according to the water-filling allocation over $\frac{\sigma^2}{\lambda_{2,r}^2}$ and to choose

$$\mathbf{G} = \mathbf{V}_2(\sigma^2\mathbf{I} + \mathbf{\Lambda}_A^2)^{-\frac{1}{2}}\mathbf{D}_p^{\frac{1}{2}}\mathbf{U}_A^H, \quad (2.18)$$

where \mathbf{D}_p is a diagonal matrix whose diagonal elements are the powers p_r . Notice that the relay still has to know the SVD of the matrix \mathbf{A} , even if the water-filling is performed directly over the modes of the second-hop channel \mathbf{H}_2 .

2.4.2 Constant-power water-filling

In the previous Section we showed how a water-filling allocation over the modes of \mathbf{H}_2 is an approximation of the exact solution to (2.7). Unfortunately, exact water-filling is *not* a computationally efficient algorithm. Therefore, various low-complexity algorithms have been recently proposed [33] [7], most of them based on *constant-power water-filling*. The concept underlying constant-power water-filling algorithms is very simple. The capacity of N parallel Gaussian channels is simply the sum of the capacity on each sub-channel,

$$C = \sum_{i=1}^N \log\left(1 + \frac{p_i}{\sigma_i^2}\right). \quad (2.19)$$

Since each element of the sum in (2.19) is a logarithmic function of power, it is rather insensitive to *exact* power allocation, except when σ_i is high (i.e., low SNR conditions). So the crucial point in water-filling is to allocate the correct amount of power to *low* SNR sub-channels. This motivates the the search for simpler power allocation scheme.

In a constant-power allocation algorithm, the power at the transmitter disposal is *equally* subdivided among the first n best sub-channels. What determines the complexity of the algorithm is the procedure to find the number of sub-channels to use, n . A simple strategy would be to start with one sub-channel and then in each step compute the rate using one more sub-channel, until the rate decreases. This

would be similar to the scheme proposed in [33], where zero power was allocated to sub-channels that would receive zero-power in exact water-filling, and constant power was allocated to sub-channels that would receive positive power in exact water-filling. However, such algorithms would require to compute a lot of logarithmic operations.

Recently, a log-free constant-power water-filling algorithm has been proposed in [29]. In this algorithm, the number of used sub-channels n is decided with a very simple procedure, which requires only a scalar division in each step. In the following, we will use this algorithm to perform a constant-power water-filling over the modes of \mathbf{H}_2 . The algorithm can be detailed for our case as

```

initialize  $m = 1$ ;
repeat
(1) compute  $p_0 = \frac{P_R}{m}$  ;
(2) if  $p_0 + \frac{\sigma^2}{\lambda_{2,1}^2} \leq \frac{\sigma^2}{\lambda_{2,m+1}^2}$  set  $n = m$  and break;
else set  $m = m + 1$  ;

```

So in each step only a single division is performed and the complexity is very low. The power allocated to the r -th sub-channel will be

$$p_r = \begin{cases} p_0 & \text{if } r \leq n \\ 0 & \text{if } r > n \end{cases} \quad (2.20)$$

The algorithm is graphically illustrated in fig. 2.6. In our case the output of the algorithm is the power allocated to the r -th sub-channel p_r . To compute the multiplicative coefficient g_r it is necessary to further perform at most N divisions and square roots, since

$$g_r = \sqrt{\frac{p_r}{\lambda_{A,r}^2 + \sigma^2}} \quad (2.21)$$

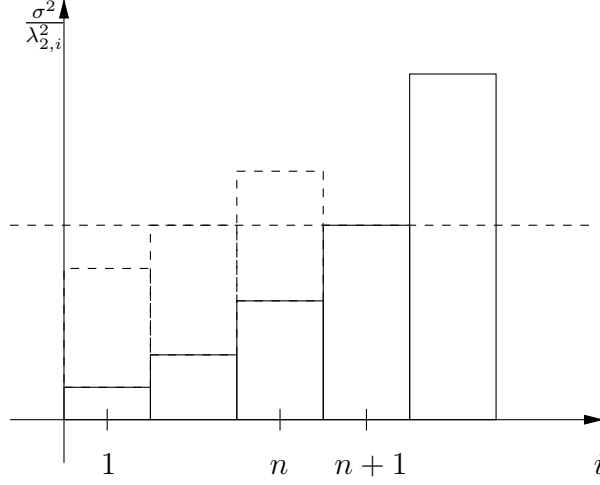


Figure 2.6: The constant-power water-filling algorithm.

2.4.3 A duality gap bound on the accuracy of constant-power water-filling

In Sec. 2.4.1 we found that the exact solution of (2.7) can be approximated with a water-filling allocation over the modes of \mathbf{H}_2 . In Sec. 2.4.2 we proposed a constant-power water-filling allocation to further reduce the computational complexity. In this Section we will develop an analytical bound on the error of a generic approximate solution to (2.7). In fact, since the original problem (2.6) is convex, it is possible to compute a duality gap bound on the performance of sub-optimal algorithms.

The result presented in Sec. A.1 on the duality gap is applied here to our particular water-filling problem. In the following derivation we will follow the guidelines found in [29], where the duality gap for approximate single-user water-filling was derived. The Lagrangian of the optimization problem (2.7) is

$$L(|g_1|^2, \dots, |g_N|^2, \boldsymbol{\gamma}, \nu) = - \sum_{r=1}^N \log \left(1 + \frac{\lambda_{A,r}^2 \lambda_{2,r}^2 |g_r|^2}{\sigma^2 + \sigma^2 \lambda_{2,r}^2 |g_r|^2} \right) - \sum_{r=1}^N \gamma_r |g_r|^2 + \nu \left(\sum_{r=1}^N (\lambda_{A,r}^2 + \sigma^2) |g_r|^2 - P_R \right). \quad (2.22)$$

At the infimum, the derivative² of the Lagrangian with respect to $|g_r|^2$ must be zero

$$\frac{\partial L}{\partial |g_r|^2} = -\frac{\lambda_{A,r}^2 \lambda_{2,r}^2 \sigma^2}{(\sigma^2 + \sigma^2 \lambda_{2,r}^2 |g_r|^2)^2 \left(1 + \frac{\lambda_{A,r}^2 \lambda_{2,r}^2 |g_r|^2}{\sigma^2 + \sigma^2 \lambda_{2,r}^2 |g_r|^2}\right)} - \gamma_r + \nu(\lambda_{A,r}^2 + \sigma^2) |g_r|^2 = 0, \quad (2.23)$$

thus yielding the water-filling condition

$$p_r + \frac{\sigma^2}{\lambda_{2,r}^2} = \sqrt{\left(\frac{\eta_r}{2}\right)^2 + \eta_r \frac{1}{\nu - \gamma_r / (\lambda_{A,r}^2 + \sigma^2)}} - \frac{\eta_r}{2}, \quad (2.24)$$

where we defined $p_r = (\lambda_{A,r}^2 + \sigma^2) |g_r|^2$ and $\eta_r = \frac{\lambda_{A,r}^2}{\lambda_{2,r}^2}$. Substituting (2.24) in (2.22) and calculating the duality gap we obtain

$$\Gamma = \nu P_R - \sum_{r=1}^N \frac{\eta_r}{p_r + \frac{\sigma^2}{\lambda_{2,r}^2} + \eta_r} \frac{p_r}{p_r + \frac{\sigma^2}{\lambda_{2,r}^2}} \quad (2.25)$$

To express the gap exclusively in primal variables, a suitable ν needs to be found. Since

$$\nu - \frac{\gamma_r}{\lambda_{A,r}^2 + \sigma^2} = \frac{\eta_r}{p_r + \frac{\sigma^2}{\lambda_{2,r}^2} + \eta_r} \frac{1}{p_r + \frac{\sigma^2}{\lambda_{2,r}^2}} \quad (2.26)$$

and ν and γ_r need to be positive, the smallest non-negative ν is then

$$\nu = \max_r \left\{ \frac{\eta_r}{p_r + \frac{\sigma^2}{\lambda_{2,r}^2} + \eta_r} \frac{1}{p_r + \frac{\sigma^2}{\lambda_{2,r}^2}} \right\} = \frac{1}{\min_r \left\{ \left(\frac{p_r + \sigma^2 / \lambda_{2,r}^2}{\eta_r} + 1 \right) \left(p_r + \frac{\sigma^2}{\lambda_{2,r}^2} \right) \right\}}. \quad (2.27)$$

²For the sake of the simplicity of the derivation, here we assume that the logarithm is in natural basis.

Since $P_R = \sum_{r=1}^N p_r$ the final expression for Γ reads³

$$\Gamma = \sum_{r=1}^N \left[\frac{p_r}{\min_j \left\{ \left(\frac{p_j + \sigma^2 / \lambda_{2,j}^2}{\eta_j} + 1 \right) \left(p_j + \frac{\sigma^2}{\lambda_{2,j}^2} \right) \right\}} - \frac{\eta_r}{p_r + \frac{\sigma^2}{\lambda_{2,r}^2}} \frac{p_r}{p_r + \frac{\sigma^2}{\lambda_{2,r}^2}} \right] \quad (2.28)$$

A sub-optimal solution to the problem (2.7) is then *at most* Γ nats/s/Hz away from the optimal water-filling solution (2.8). From (2.28) it is clear that our choice of ν (2.27) is good, since if the primal feasible $|g|_r^2$ are optimal, then $\Gamma = 0$ as expected. Note that, coherently with the results in [29], if we let $\eta_r \rightarrow +\infty$, (2.28) reduces to the expression of the single-user water-filling duality gap

$$\Gamma_{su} = \sum_{r=1}^N \left[\frac{p_r}{\min_j \left\{ \left(p_j + \frac{\sigma^2}{\lambda_{2,j}^2} \right) \right\}} - \frac{p_r}{p_r + \frac{\sigma^2}{\lambda_{2,r}^2}} \right]. \quad (2.29)$$

2.4.4 Low-complexity iterative solution

Building on the results of the previous Sections, a low-complexity version of the iterative algorithm presented in Sec. 2.3.3 can be devised. In the iterative algorithm of Sec. 2.3.3 two optimizations have to be performed in each iteration: a single-user water-filling to solve (2.5) and the non-linear water-filling (2.8) previously analyzed to solve (2.7). As shown in Sec. 2.4.2, the non-linear water-filling (2.8) can be approximated using a constant-power allocation algorithm. Of course, as detailed in [29], the same algorithm can be used to calculate a sub-optimal constant-power solution to (2.5). The low-complexity iterative algorithm can be detailed as

```

initialize  $\mathbf{Q} = \mathbf{0}$ ,  $\mathbf{G} = \mathbf{I}$ ;
repeat
(a) use the constant-power allocation algorithm in [29] to find a
sub-optimal solution to (2.5) for  $\mathbf{Q}$  keeping  $\mathbf{G}$  fixed;
(b) use the constant-power allocation algorithm presented in Sec. 2.4.2

```

³This form of Γ is in nats/s/Hz. To compute the duality gap in bit/s/Hz (2.28) has to be divided by $\ln 2$.

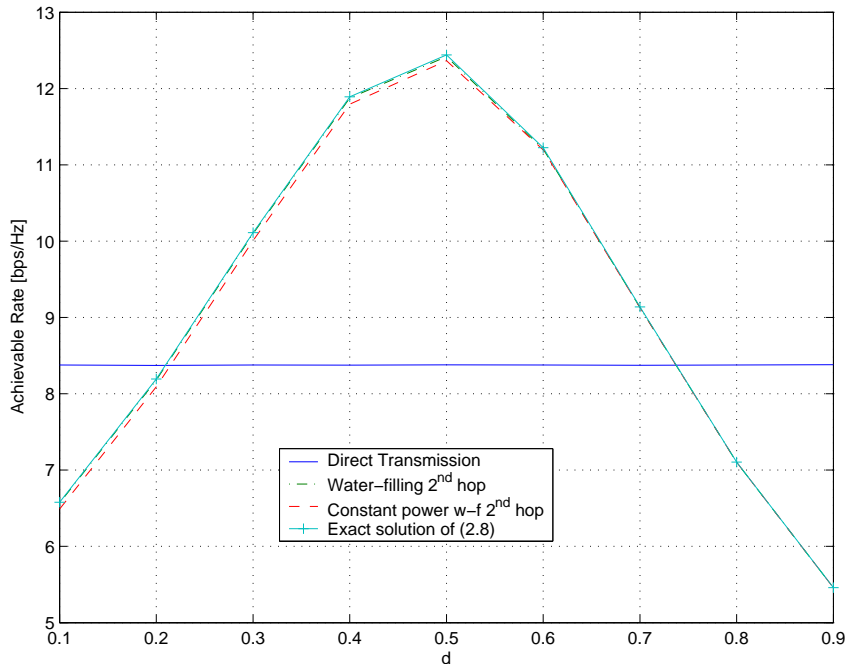


Figure 2.7: Achievable rates of the two sub-optimal algorithms and exact (2.8) water-filling, $\mathbf{Q} = N^{-1}\mathbf{I}$, $N = 10$, $\frac{1}{\sigma^2} = 0dB$.

to find a sub-optimal solution to (2.6) for \mathbf{G} keeping \mathbf{Q} fixed; until the rate (2.3) converges.

It is not possible to prove absolute convergence for this algorithm. Nevertheless, as the algorithm presented in Sec. 2.3.3, if the problem (2.2) is not ill-conditioned, it has shown to rapidly converge⁴ to a unique solution.

2.4.5 Simulation results

In this Section we present simulation results on the performances of the low-complexity algorithms proposed. First, we analyze the correctness of our intuitions developed in Sec. 2.4.1 and 2.4.2 about the approximation of the solution of (2.8). Here the source input covariance matrix is assumed isotropic $\mathbf{Q} = N^{-1}\mathbf{I}$, the relay power

⁴around 4-5 iteration with a tolerance of 10^{-2} on the rate.

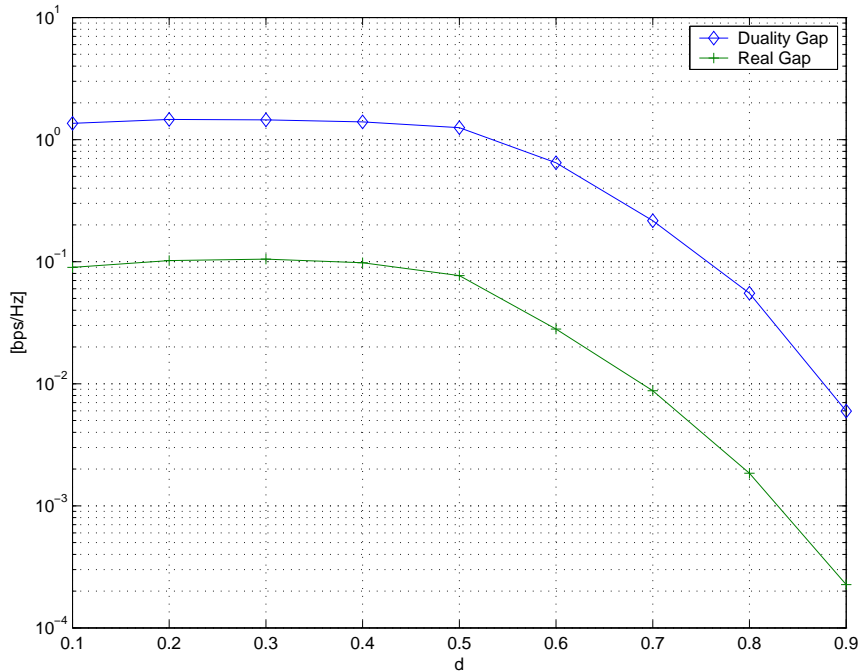


Figure 2.8: Duality gap and effective gap for the constant-power allocation algorithm described in Sec. 2.4.2, $\mathbf{Q} = N^{-1}\mathbf{I}$, $N = 10$, $\frac{1}{\sigma^2} = 0dB$.

constraint $P_R = 1$, the number of antennas $N = 10$, and the SNR $\frac{1}{\sigma^2} = 0dB$. Fig. 2.7 shows the ergodic achievable rates of the exact solution of (2.8), of the water-filling allocation over the second-hop modes (as described in Sec. 2.4.1) and of the constant-power water-filling allocation over the second-hop modes (as described in Sec. 2.4.2). It is clear that both of the two approximate solutions (the water-filling and the constant-power water-filling) are very close to the optimum.

The duality and effective gap for the constant-power water-filling allocation is plotted in fig. 2.8. The real gap is much tighter than the analytical bound, as expected.

In fig. 2.9 we can see how the average number of channels n used by the relay according to the constant-power allocation algorithm grows as the relay approaches the destination. This was as well expected since for $d \rightarrow 1$ the channel \mathbf{H}_2 becomes stronger according to our geometrical model.

Finally, we analyze the performances of the low-complexity iterative algorithm

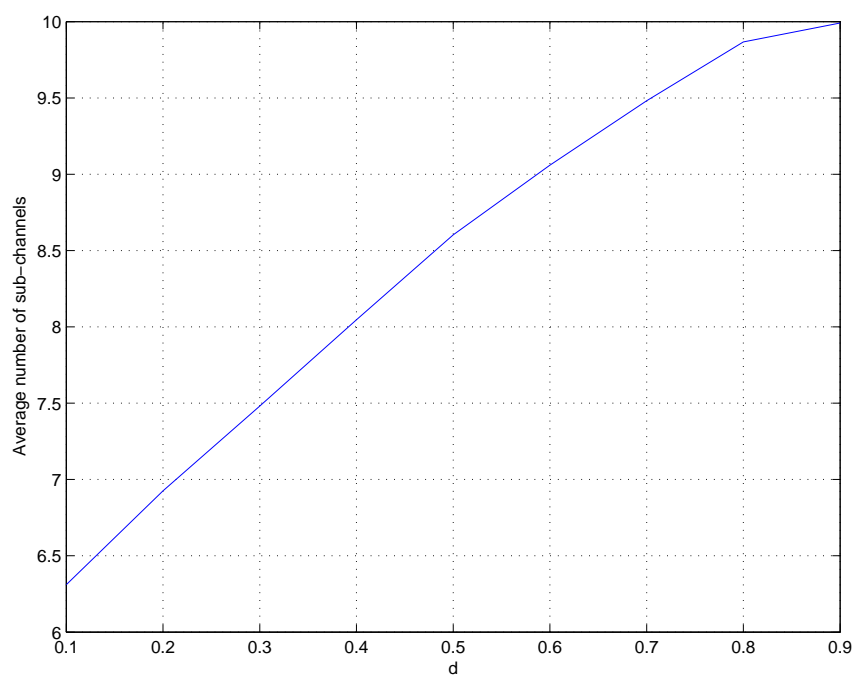


Figure 2.9: Number of sub-channels n used according to the constant-power allocation algorithm, $\mathbf{Q} = N^{-1}\mathbf{I}$, $N = 10$, $\frac{1}{\sigma^2} = 0dB$.

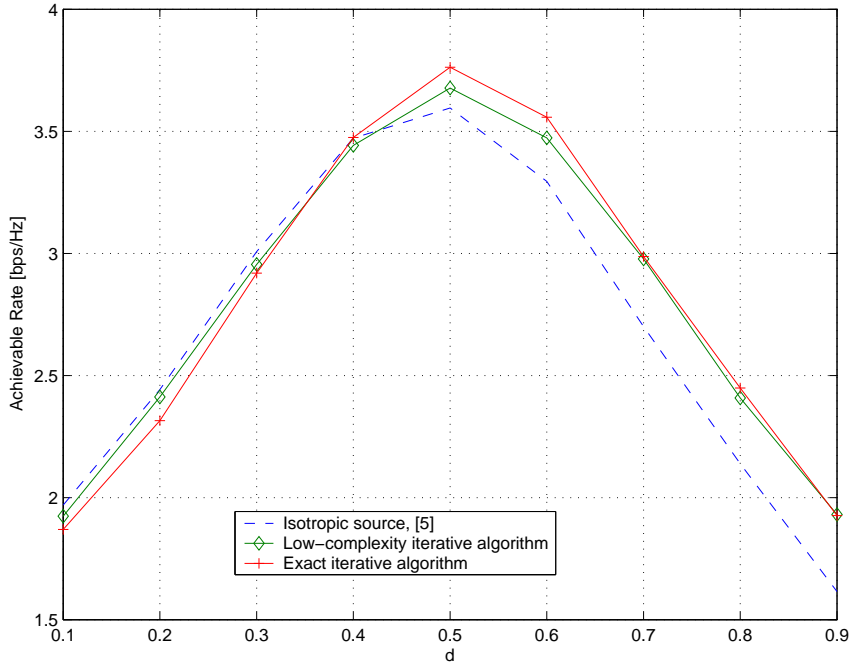


Figure 2.10: Achievable rate of the low-complexity iterative algorithm compared with the exact iterative algorithm presented in Sec. 2.3.3 and the communication scheme of [5], $N = 3$, $\frac{1}{\sigma^2} = 0dB$.

developed in Sec. 2.4.4 to approximate the exact solution of the iterative algorithm of Sec. 2.3. The ergodic achievable rate of the low-complexity iterative algorithm is shown in fig. 2.10 and compared with the achievable rates of the exact algorithm and the communication scheme of [5], for $N = 3$, $\frac{1}{\sigma^2} = 0dB$ and $P = 1$, $P_R = 1$. It is seen that the rate of the low-complexity algorithm is close to the rate of the exact algorithm for a wide range of values of d . Further, for $d < 0.35$ the algorithm proposed here even exceeds the rate of the exact algorithm, and its rate is close to that of the scheme of [5]. In fact, we have to keep in mind that even the exact iterative algorithm is just a sub-optimal feasible solution to the general problem (2.2).

In fig. 2.11 the number of sub-channels used by the source and by the relay on average with respect to channels realizations is shown for the low-complexity iterative technique presented in Sec. 2.4.4. The fact that the source and the relay allocate power to a different number of channels is not surprising. In fact, this is due to the

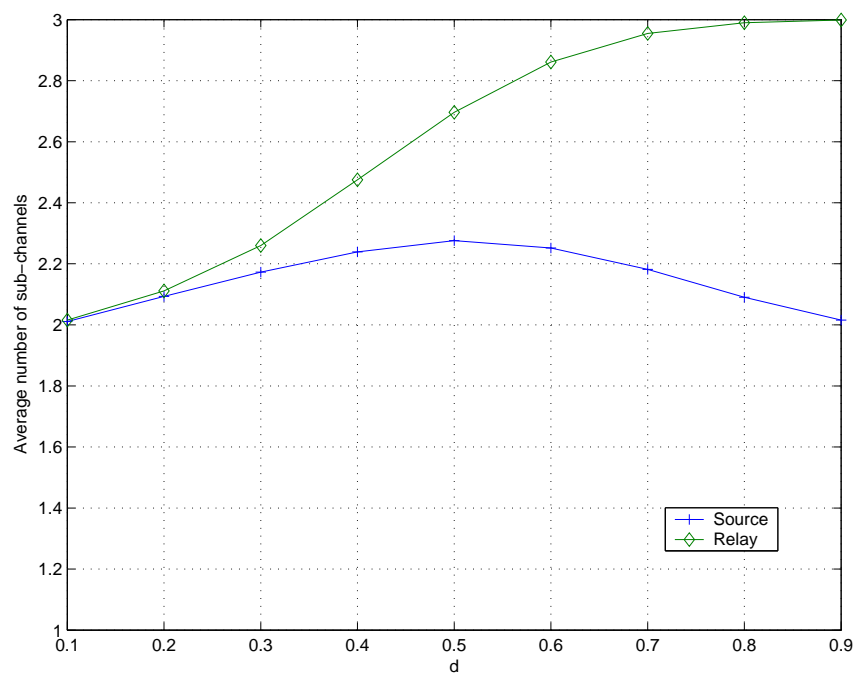


Figure 2.11: Number of sub-channels used by the relay and source, according to the low-complexity iterative algorithm, $N = 3$, $\frac{1}{\sigma^2} = 0dB$.

iterative approach we have used in Sec. 2.3 to solve the problem (2.2). In each step, \mathbf{Q} (\mathbf{G}) is chosen depending on the eigenmodes of an equivalent channel that depends on the value assigned to \mathbf{G} (\mathbf{Q}) in the previous iteration. This means that the sub-channels used by the source for transmission are not the same sub-channels used by the relay for reception. In AF systems the relay node does not have to decode, and so it does not matter if it cannot reconstruct the streams transmitted by the source. Moreover, the destination has to have full knowledge of these operations in order to decode the message from the source. Indeed, expanding the first- and second-hop channel matrices with the SVD as $\mathbf{H}_1 = \mathbf{U}_1 \mathbf{\Lambda}_1 \mathbf{V}_1^H$ and $\mathbf{H}_2 = \mathbf{U}_2 \mathbf{\Lambda}_2 \mathbf{V}_2^H$, the relay channel could be easily converted into N parallel SISO AF links choosing

$$\mathbf{Q} = \mathbf{V}_1 \mathbf{D}_q \mathbf{V}_1^H \quad (2.30)$$

$$\mathbf{G} = \mathbf{V}_2 \mathbf{D}_g \mathbf{U}_1^H. \quad (2.31)$$

However, it will be clear in the next chapters that this choice does not lead to any suggestion for a solution of the optimization of the collaborative AF relay case.

One-iteration performance

It is of great interest for the practical implementation of the low-complexity algorithm previously presented to analyze its performance after just one step of the iterative procedure. In fact, performing just one iteration, only a total of three SVDs will be required: one to find the signaling sub-space for \mathbf{Q} , and two to find the sub-spaces used by the relay gain matrix \mathbf{G} . The average achievable rate after one iteration has been plotted in fig. 2.12 versus the achievable rate at convergence for $N = 3$, $\frac{1}{\sigma^2} = 0dB$. For $d < 0.6$ only a fraction of bit/s/Hz is lost, while for $d > 0.6$ we have immediate convergence. Further, in fig. 2.13 it is seen that, after just one iteration, the low-complexity iterative algorithm performs better than the scheme in [5] for $d \geq 0.5$.

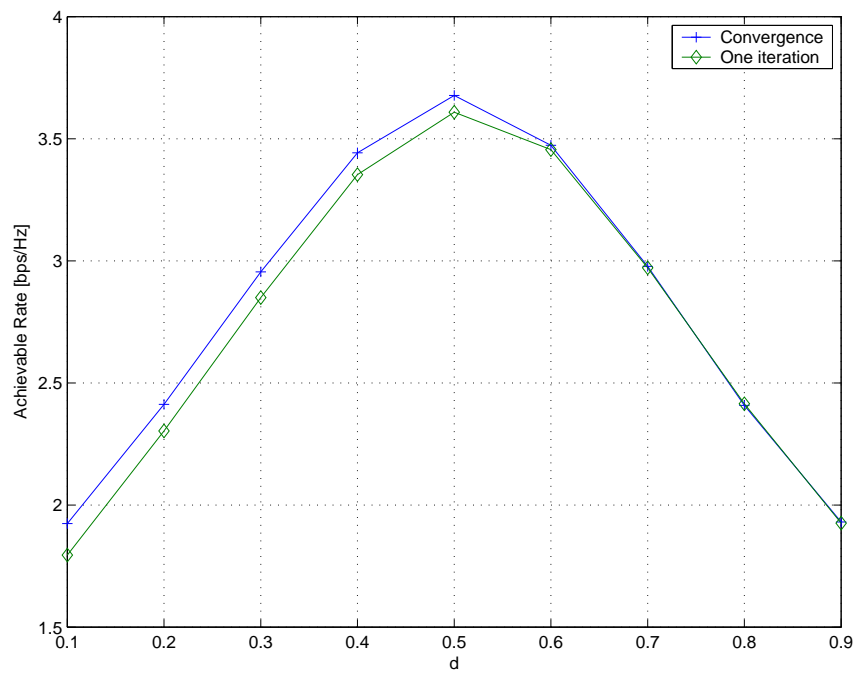


Figure 2.12: Achievable rate of the low-complexity iterative algorithm at convergence and after one iteration, $N = 3$, $\frac{1}{\sigma^2} = 0dB$.

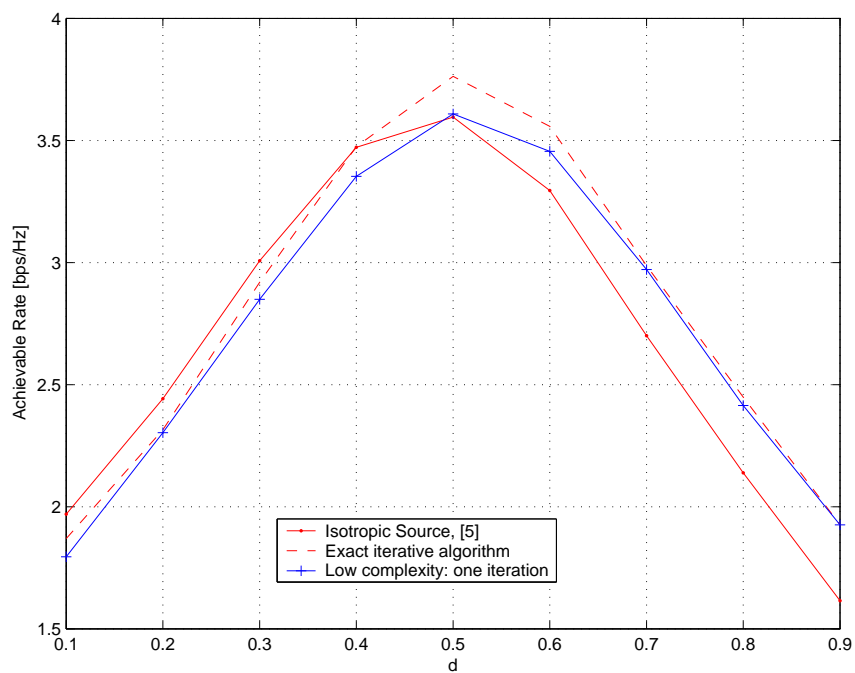


Figure 2.13: Achievable rate of the low-complexity iterative algorithm after one iteration compared to the scheme in [5] and to the exact iterative algorithm, $N = 3$, $\frac{1}{\sigma^2} = 0dB$.

Chapter 3

The AF Relay Channel with Multiple Antennas at the Relay

The main difference between *cooperative* schemes and conventional *multi-hop* relaying schemes is that the destination decodes the source message from the signals received from *both* the source and the relay node. In this chapter, a cooperative AF relay channel with multiple antennas deployed at the relay is analyzed (see fig. 3.1). For this system the optimal relay linear processing matrix and power allocation are derived.

In [22] and [6] fixed power allocation between source and relay is assumed. Recently, power allocation strategies have been investigated for the fading relay channel by various authors. In [23] resource allocation strategies minimizing the energy-per-information-bit are presented for different protocols under an instantaneous total power constraint over the two time-slots. A rate-maximizing power allocation algorithm has been developed in [24] for an adaptive version of the DF protocol, with separate average power constraints for the source and the relay node. Coded protocols were also considered in [25], with an average total power constraint over the two time-slots. Finally, in [26] optimal power control for the minimization of the outage probability has been studied for decode- and estimate-and-forward protocols with an average sum-power constraint for the source and the relay. It should be emphasized that in these previous works all the nodes in the system were deployed with only one antenna.

In this Chapter, we consider the problem of resource allocation for a fading relay channel operating according to the AF protocol, where each node has full CSI and multiple-antennas are deployed at the relay node. The achievable rate is maximized over the linear processing at the relay and the power allocation between the source and the relay. An instantaneous sum-power constraint is imposed, such that in each time-slot the total power used by the active node(s) in the system must be equal. Notice that, in principle, a larger rate could be achieved by allowing a different total power for transmissions during the first and second time-slot [27]. However, this benefit would come at the expenses of an increased transmitted power dynamics and it will not be further investigated in this work.

The outline of the chapter is the following: in Sec. 3.1 the system model and the optimization problem are defined; in Sec. 3.2 the optimal relay linear processing is derived, whereas the optimization of the power allocation is performed in Sec 3.3. Finally, simulations results are presented in Sec 3.4.

3.1 System model and problem definition

The system under analysis, illustrated in fig. 3.1, consists of a source node, a destination node, and a multi-antenna relay node deployed with N antennas, which amplifies the received signal and forwards it towards the destination node. In order to comply with the practical half-duplex constraint, the relay transmission occurs according to a time division duplex (TDD) protocol. In the first time slot the source broadcasts the signal x_1 to the relay and the destination. In the second time slot, the relay re-transmits an amplified version of x_1 , while the source sends a second signal x_2 , chosen independently from x_1 . The time durations of the two time-slots are equal. The destination then *jointly* decodes x_1 and x_2 , given the symbols received in the two time-slots.

All the fading channels between pair of antennas are assumed to be affected by independent Rayleigh flat fading processes and additive white gaussian noise (AWGN). We also assume that every node in the system knows the state of the channels $\mathcal{H} = \{h_0, \mathbf{h}_1, \mathbf{h}_2\}$, where h_0 is the scalar source-destination channel, \mathbf{h}_1 is the $N \times 1$

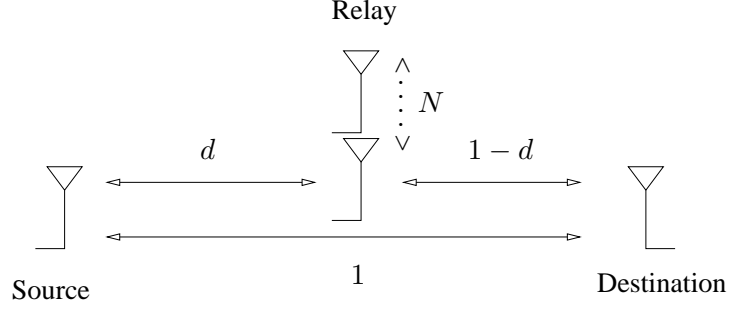


Figure 3.1: The Relay channel with multiple-antennas at the relay node.

vector source-relay channel and \mathbf{h}_2 is the $N \times 1$ vector relay-destination channel. The state \mathcal{H} is a stationary ergodic process, and it is constant over the two time-slots. The signal received by the destination in the first time-slot can be written as

$$y_1 = h_0 x_1 + n_1, \quad (3.1)$$

where x_1 is the symbol transmitted by the source in the first time-slot and n_1 is the noise sample at the destination, assumed to have distribution $\mathcal{CN}(0, N_0)$. At the same time, the relay receives a signal

$$\mathbf{y}_R = \mathbf{h}_1 x_1 + \mathbf{n}_R, \quad (3.2)$$

where \mathbf{n}_R is the noise vector at the relay, with distribution $\mathcal{CN}(\mathbf{0}, N_0 \mathbf{I}_N)$. In the second time-slot, both the source and the relay are active. The relay processes the vector signal previously received by multiplication with a $N \times N$ matrix \mathbf{G} . Thus, the symbol transmitted by the relay can be expressed as

$$\mathbf{x}_R = \mathbf{G} \mathbf{y}_R = \mathbf{G} \mathbf{h}_1 x_1 + \mathbf{G} \mathbf{n}_R. \quad (3.3)$$

The overall signal received by the destination in the second time-slot is

$$y_2 = \mathbf{h}_2^H \mathbf{G} \mathbf{h}_1 x_1 + h_0 x_2 + \mathbf{h}_2^H \mathbf{G} \mathbf{n}_R + n_2, \quad (3.4)$$

where x_2 is the signal transmitted by the source and n_2 is the noise sample at the

destination, assumed to have distribution $\mathcal{CN}(0, N_0)$. Finally, we can more compactly express the vector input/output relation for this channel as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} h_0 & 0 \\ \mathbf{h}_2^H \mathbf{G} \mathbf{h}_1 & h_0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 & \mathbf{0}^H & 0 \\ 0 & \mathbf{h}_2^H \mathbf{G} & 1 \end{bmatrix} \begin{bmatrix} n_1 \\ \mathbf{n}_R \\ n_2 \end{bmatrix}. \quad (3.5)$$

Note that, since x_1 and x_2 are assumed to be independent, the model (3.5) corresponds to a two-user multiple access channel with two antennas at the receiver, and thus successive decoding is a capacity-achieving decoding strategy [3].

Based on the channel state information \mathcal{H} , the source node allocates the power $P_{s_1}(\mathcal{H})$ and $P_{s_2}(\mathcal{H})$ for transmission during the first and second time-slot respectively, while the relay node employs the power allocation policy $P_r(\mathcal{H})$. This later assumption affects the design of the matrix $\mathbf{G}(\mathcal{H})$ since the power transmitted by the relay is

$$P_r(\mathcal{H}) = \text{tr}(\mathbf{x}_R \mathbf{x}_R^H) = \text{tr}(\mathbf{G}(\mathcal{H}) \mathbf{h}_1 P_{s_1}(\mathcal{H}) \mathbf{h}_1^H \mathbf{G}^H(\mathcal{H}) + \mathbf{G}(\mathcal{H}) N_0 \mathbf{G}^H(\mathcal{H})). \quad (3.6)$$

Power allocation and relay processing are jointly optimized so as to maximize the *instantaneous* achievable rate of the protocol

$$R = \frac{1}{2} I(x_1, x_2; y_1, y_2), \quad (3.7)$$

where $I(x_1, x_2; y_1, y_2)$ is the mutual information between the source input (x_1, x_2) and the output at the destination (y_1, y_2) , and the factor $1/2$ accounts for the time-division operation. An instantaneous power constraint on each time-slot is enforced so that

the problem can be stated as

$$\begin{aligned} & \max_{\mathbf{G}(\mathcal{H}), \theta(\mathcal{H})} R & (3.8) \\ \text{s.t.} & \begin{cases} P_{s_1}(\mathcal{H}) = 1 \\ P_{s_2}(\mathcal{H}) = \theta(\mathcal{H}) \\ P_r(\mathcal{H}) = 1 - \theta(\mathcal{H}) \end{cases} \end{aligned}$$

$\forall \mathcal{H}$, where $\theta(\mathcal{H}) \in [0, 1]$.

The power constraints in (3.8) allow the transmission scheme to encompass and generalize direct transmission and the AF cooperative protocols in [22] and [6]. In fact, if $\theta(\mathcal{H}) = 1$ the relay is not used at all, neither during the first nor in the second time-slot, and the communication scheme boils down to direct transmission from the source to the relay node; if $\theta(\mathcal{H}) = 0$ only the relay is active during the second time-slot and the communication protocol is the same as in [22]; finally, if $\theta(\mathcal{H}) = 1/2$ the communication protocol employs both the relay and the source in the second time slot, and thus resembles the scheme introduced in [6]. Notice that, under the considered ergodicity assumption, we could have also enforced a long-term power constraint in (3.8): this modification would complicate the analysis and is not further considered in this work.

In order to tackle the optimization problem (3.8), is convenient to expand the instantaneous achievable rate (3.7) by using the chain rule for the mutual information [3]

$$R = \underbrace{\frac{1}{2}I(x_1; y_1)}_{I_1} + \underbrace{\frac{1}{2}I(x_1, x_2; y_2|y_1)}_{I_2}, \quad (3.9)$$

where we have used the fact that $I(x_2; y_1|x_1) = 0$. In (3.9) I_1 accounts for the source transmission during the first time-slot, while I_2 measures the contribution of the relay and source transmissions during the second time-slot. Given (3.5) and (3.9), for any power allocation policies $P_{s_1}(\mathcal{H})$, $P_{s_2}(\mathcal{H})$, $P_r(\mathcal{H})$ and linear processing at the relay

$\mathbf{G}(\mathcal{H})$, the achievable rate components read

$$I_1 = \frac{1}{2}C \left(\frac{|h_0|^2 P_{s_1}(\mathcal{H})}{N_0} \right) \quad (3.10)$$

$$I_2 = \frac{1}{2}C \left(\frac{|h_0|^2 P_{s_2}(\mathcal{H}) + \frac{\mathbf{h}_2^H \mathbf{G}(\mathcal{H}) \mathbf{h}_1 P_{s_1}(\mathcal{H}) \mathbf{h}_1^H \mathbf{G}^H(\mathcal{H}) \mathbf{h}_2}{1 + \frac{|h_0|^2 P_{s_1}(\mathcal{H})}{N_0}}}{N_0(1 + \mathbf{h}_2^H \mathbf{G}(\mathcal{H}) \mathbf{G}^H(\mathcal{H}) \mathbf{h}_2)} \right), \quad (3.11)$$

where we have defined $C(x) := \log(1 + x)$. From (3.8), (3.9), (3.10), (3.11) it is clear that the optimal power allocation requires $P_{s_1}(\mathcal{H}) = 1, \forall \mathcal{H}$. On the other hand, the optimization of powers $P_{s_2}(\mathcal{H})$ and $P_r(\mathcal{H})$, and linear processing $\mathbf{G}(\mathcal{H})$ boils down to the maximization of the term I_2 (3.11). In the next Sections, we first discuss the optimization of the relay linear processing $\mathbf{G}(\mathcal{H})$ (Sec. 3.2), and the of power policies $P_{s_2}(\mathcal{H})$ and $P_r(\mathcal{H})$ (Sec. 3.3).

3.2 Optimization of the relay linear processing

In this Section, the expression for the optimal linear processing matrix at the relay $\mathbf{G}(\mathcal{H})$ is derived. From (3.11) it is easy to prove that the optimal $\mathbf{G}(\mathcal{H})$ can be written as the outer product of the beamformers for the channels \mathbf{h}_1 and \mathbf{h}_2

$$\mathbf{G} = \frac{g \mathbf{h}_2 \mathbf{h}_1^H}{\|\mathbf{h}_2\| \|\mathbf{h}_1\|}, \quad (3.12)$$

where the scalar normalization factor g in (3.12) is determined by the relay power constraint (3.6). This can be restated, using the constraints in (3.8) and (3.12) as

$$P_r(\mathcal{H}) = \|\mathbf{h}_1\|^2 |g|^2 + N_0 |g|^2 = 1 - \theta(\mathcal{H}), \quad (3.13)$$

which implies the condition

$$g = \sqrt{\frac{1 - \theta(\mathcal{H})}{N_0 + \|\mathbf{h}_1\|^2}}. \quad (3.14)$$

Substituting the optimal expressions (3.12) and (3.14) into the expression of I_2 (3.11), we get

$$I_2 = \frac{1}{2}C \left(\frac{|h_0|^2\theta(\mathcal{H}) + \frac{\|\mathbf{h}_2\|^2\|\mathbf{h}_1\|^2}{1+\frac{|h_0|^2}{N_0}} \frac{1-\theta(\mathcal{H})}{N_0+\|\mathbf{h}_1\|^2}}{N_0(1 + \|\mathbf{h}_2\|^2 \frac{1-\theta(\mathcal{H})}{N_0+\|\mathbf{h}_1\|^2})} \right), \quad (3.15)$$

which depends only on the power allocation policy $\theta(\mathcal{H})$.

3.3 Optimization of the power allocation

In this section, the optimal power allocation policy is derived. As discussed in the previous Section, this problem boils down to the maximization of I_2 in (3.15) over the fraction of power allocated to the source in the second time-slot $\theta(\mathcal{H})$. Therefore the problem can be stated as

$$\max_{\theta \in [0,1]} f_0(\theta) \quad (3.16)$$

with

$$f_0(\theta) = C \left(\frac{|h_0|^2\theta + \frac{\|\mathbf{h}_2\|^2\|\mathbf{h}_1\|^2}{1+\frac{|h_0|^2}{N_0}} \frac{1-\theta}{N_0+\|\mathbf{h}_1\|^2}}{N_0(1 + \|\mathbf{h}_2\|^2 \frac{1-\theta}{N_0+\|\mathbf{h}_1\|^2})} \right).$$

After tedious calculations, it can be verified that

$$\frac{df_0(\theta)}{d\theta} = \frac{\alpha}{\beta(\theta)}, \quad (3.17)$$

where

$$\alpha = -(N_0 + \|\mathbf{h}_1\|^2) [\|\mathbf{h}_1\|^2 \|\mathbf{h}_2\|^2 N_0 - |h_0|^2(N_0 + |h_0|^2)(N_0 + \|\mathbf{h}_1\|^2 + \|\mathbf{h}_2\|^2)], \quad (3.18)$$

and $\beta(\theta)$ has a cumbersome expression which is positive for every value of $\theta \in [0, 1]$. We can then conclude that $f_0(\theta)$ is monotone in the interval $[0, 1]$, and that the maximum is achieved in $\theta = \{0, 1\}$, according to the sign of α . We can summarize

this conclusion as

$$\begin{cases} \theta = 0, & \text{if } \alpha < 0 \\ \theta = 1, & \text{if } \alpha > 0 \end{cases}. \quad (3.19)$$

After some algebraic manipulations the optimal strategy for power sharing can be expressed as

$$\begin{cases} \theta(\mathcal{H}) = 0 & \text{if } K_D(\mathcal{H}) < K_L(\mathcal{H}) \\ \theta(\mathcal{H}) = 1 & \text{if } K_D(\mathcal{H}) > K_L(\mathcal{H}) \end{cases} \quad (3.20)$$

where $K_D(\mathcal{H}) = \frac{|h_0|^2}{N_0}$ is the argument of $C(x)$ in the second time-slot term in (3.15) if $\theta = 1$ (i.e., it corresponds to direct transmission without the use of the relay), whereas $K_L(\mathcal{H}) = \frac{\|\mathbf{h}_1\|^2 \|\mathbf{h}_2\|^2}{(1 + \frac{|h_0|^2}{N_0}) N_0 (N_0 + \|\mathbf{h}_1\|^2 + \|\mathbf{h}_2\|^2)}$ is the argument of $C(x)$ in the second time-slot term in (3.15) if $\theta = 0$, and corresponds to the AF protocol presented in [22]. In other words, for any realization of the channels it is optimal to use either direct transmission or the protocol in [22], depending upon which scheme achieves the higher rate. Under the constraint of a fixed total power for each time-slot, the scheme presented in [6] is never optimal.

As a final remark we note that if we let $N_0 \rightarrow 0$ (i.e., asymptotically with respect to the signal-to-noise ratio), it is easily seen that $K_D(\mathcal{H}) \rightarrow \infty$, while

$$\lim_{N_0 \rightarrow 0} K_L(\mathcal{H}) = \frac{\|\mathbf{h}_1\|^2 \|\mathbf{h}_2\|^2}{|h_0|^2 (\|\mathbf{h}_1\|^2 + \|\mathbf{h}_2\|^2)}. \quad (3.21)$$

Thus for high SNR $K_D > K_L$ is always verified and the asymptotic optimal allocation is $\theta = 1$, i.e., direct transmission is optimal and the AF protocol is not advantageous.

3.4 Simulations results

In order to get insight into our results, we assume that the relay is located on a line between the source and the destination, such that the average power of the channels gains depends on the source-relay distance d and the path-loss exponent γ (see fig. 3.1). Let us first consider the unfaded case (AWGN channel), where $|h_0|^2 = 1$, $\|\mathbf{h}_1\|^2 = \frac{1}{d^\gamma} N$, $\|\mathbf{h}_2\|^2 = \frac{1}{(1-d)^\gamma} N$. In this case, fig. 3.2 shows the

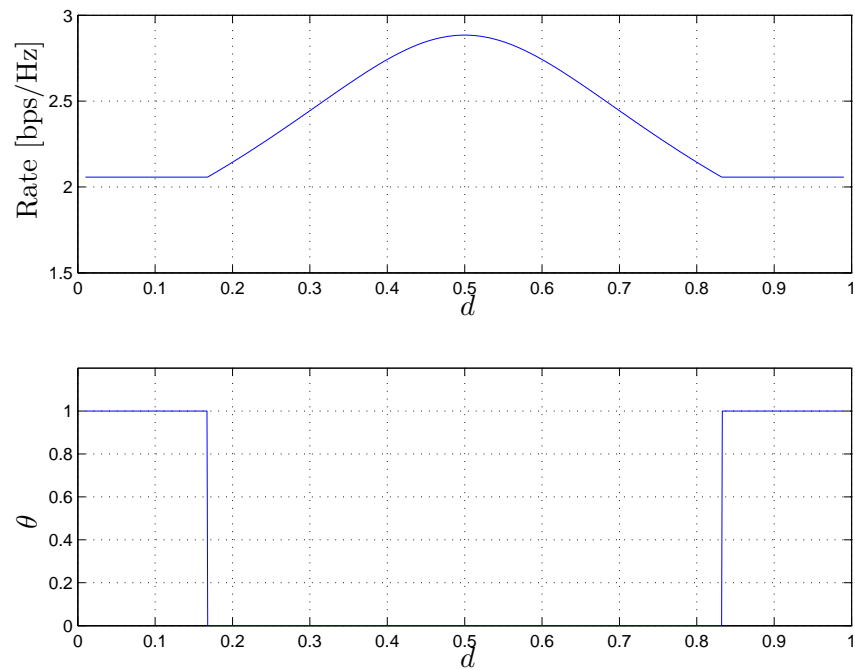


Figure 3.2: Achievable rate and optimal power allocation for the AWGN system ($N = 2$, $\gamma = 4$, $\frac{1}{N_0} = 5dB$). $\theta = 1$ implies that the relay node is silent during the second time-slot and direct transmission is employed, while $\theta = 0$ implies that the source node is silent during the second time-slot and the AF protocol in [22] is employed.

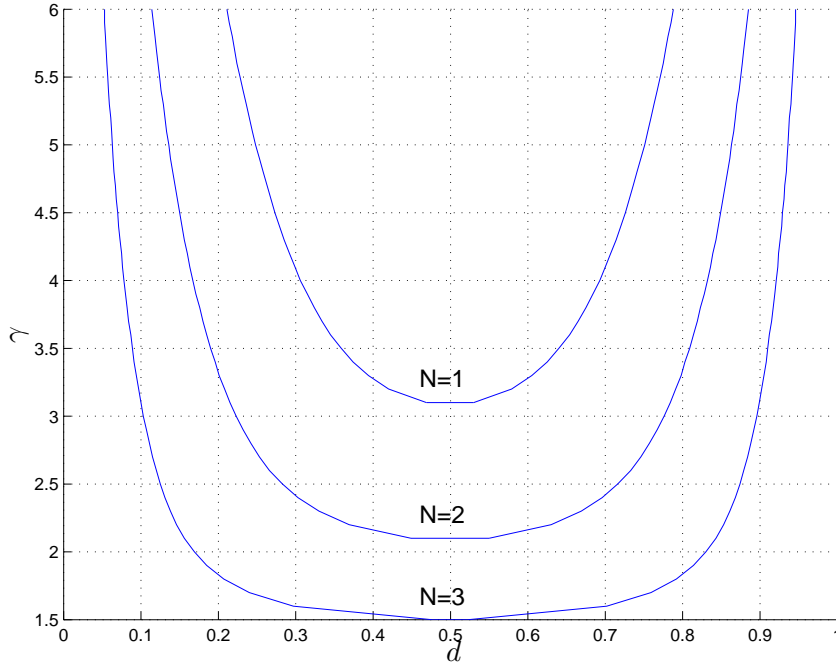


Figure 3.3: Optimality regions for different values of N ($\frac{1}{N_0} = 5dB$). In the region above each curve the AF scheme in [22] is optimal, while in the region below each curve direct transmission is optimal.

achievable rate and the optimal power allocation for $N = 2$, $\gamma = 4$, $1/N_0 = 5dB$. It is seen that the scheme in [22] is optimal when the relay is located in an interval around halfway between the source and the destination. Similar results can be obtained for different values of γ and N , showing that, increasing either the number of antennas or the path-loss exponent, the interval of d in which the relay is used grows. The regions of the $\gamma - d$ plane where either technique is optimal are plotted in fig. 3.3 for different values of N , with $1/N_0 = 5dB$. In the region above each curve $\theta = 0$ and the AF scheme in [22] is optimal, while in the region below each curve $\theta = 1$ and direct transmission is optimal. As expected, for larger number of antennas N the region in which the scheme in [22] is advantageous becomes larger.

Introducing uncorrelated Rayleigh fading, fig. 3.4 shows the ergodic achievable rate $E_{\mathcal{H}}[R]$ and the average power sharing $E_{\mathcal{H}}[\theta]$ ($E_{\mathcal{H}}[\cdot]$ denotes the expectation with respect to fading states) versus different values of the source-relay distance d for

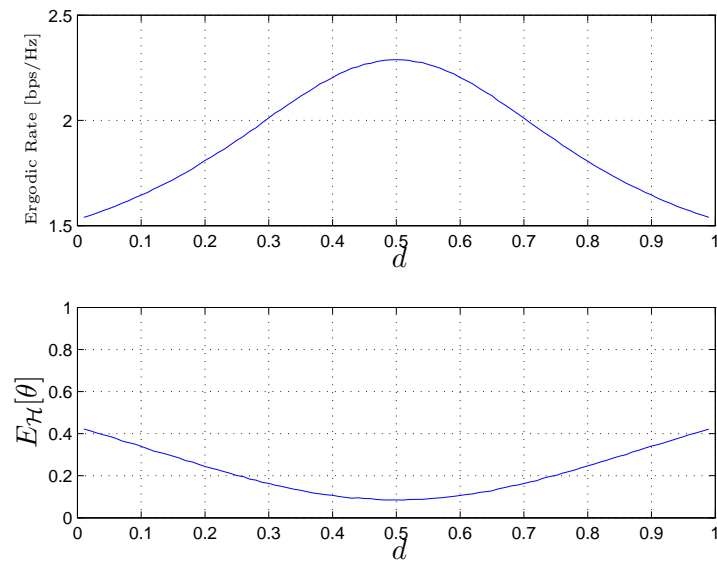


Figure 3.4: Ergodic achievable rate $E_{\mathcal{H}}[R]$ and average optimal power allocation $E_{\mathcal{H}}[\theta]$ ($N = 2$, $\gamma = 4$, $\frac{1}{N_0} = 5dB$).

$N = 2$, $\gamma = 4$, $1/N_0 = 5dB$. It is seen that $E_{\mathcal{H}}[\theta(\mathcal{H})]$ is symmetric around $d = 0.5$ and that the relay is more frequently used when it is located halfway between the source and the destination.

Chapter 4

The Cooperative MIMO AF Relay Channel

In this Chapter, we consider the case in which *all* the nodes of a cooperative AF relay channel are deployed with multiple antennas. In this case, it is not possible to derive an analytic solution for the optimal power allocation policy. However, a sub-optimal solution for the relay linear processing matrix is found using an iterative algorithm of the same fashion of the one proposed for the multi-hop channel in Chapter 2.

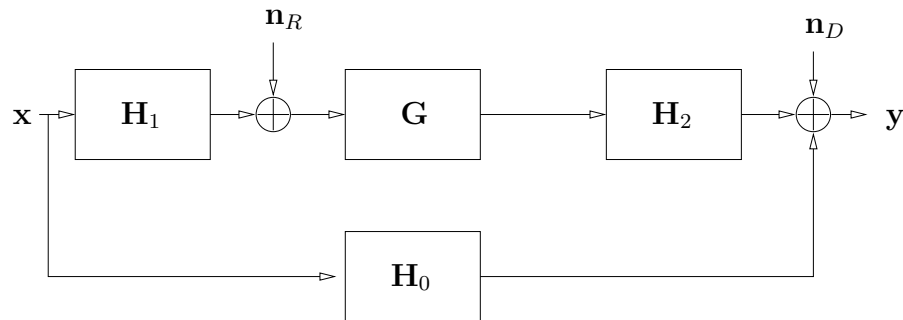


Figure 4.1: Block diagram of a cooperative MIMO Amplify-and-Forward relay system.

The multi-antenna relay channel has been recently investigated from different perspectives. The authors of [30] extended the information-theoretic results of [3] to a multi-antenna scenario and devised an algorithm to compute the input covariance

matrices that maximize the max-flow-min-cut upper bound. Performance of the AF scheme of [22] in a multi-antenna setting was analyzed in [5]. In particular, [5] derived the optimal linear processing matrix at the relay, for both multi-hop and cooperative MIMO AF relay systems, under the assumption of perfect channel state information (CSI) at each node and isotropic covariance matrix for the symbol transmitted by the source.

In this Chapter, we consider multi-hop and cooperative MIMO AF relay systems with perfect CSI at each node as in [5]. However, differently from [5]: (i) for both multi-hop and cooperative systems, the covariance matrix of the symbols transmitted by the source is *not* constrained to be isotropic; (ii) for the cooperative scenario, the considered AF protocol is the one presented in [6], whereby the source transmits in *both* time-slots (not only in the first). The problem of maximizing the achievable rate over the source covariance matrices and linear processing matrix at the relay is formulated. An iterative algorithm, capable of finding a sub-optimal feasible solution, is proposed for both multi-hop and cooperative cases, and proved by numerical simulations to outperform known schemes under broad channel conditions.

In Sec. 4.1 the system model is presented. The optimization problem we tackle is detailed in Sec. 4.2, whereas an iterative sub-optimal solution is proposed in Sec. 4.3. Finally, in Sec. 4.4, a low-complexity iterative algorithm is devised in order to approximate the exact solution of the iterative algorithm defined in Sec. 4.3.

4.1 System model

The cooperative MIMO Amplify-and-Forward relay is illustrated in fig.4.1. In the first time-slot the source transmits a first signal \mathbf{x}_1 to *both* the relay and the destination; in the second time-slot the relay retransmits \mathbf{x}_1 after a linear transformation, while the source transmits a second signal \mathbf{x}_2 , independent from \mathbf{x}_1 . At the end of the second time-slot, the destination *jointly* decodes $(\mathbf{x}_1, \mathbf{x}_2)$ from the signals received in the two time-slots.

The signal received by the destination node during the first time-slot is $\mathbf{y}_1 =$

$\mathbf{H}_0\mathbf{x}_1 + \mathbf{n}_{D,1}$, where \mathbf{x}_1 is the $N \times 1$ source input symbol, \mathbf{H}_0 is the $N \times N$ source-destination channel, and $\mathbf{n}_{D,1}$ is the $N \times 1$ noise vector at the destination. During the same time-slot, the relay node receives a signal $\mathbf{y}_R = \mathbf{H}_1\mathbf{x}_1 + \mathbf{n}_{R,1}$, where \mathbf{H}_1 is the $N \times N$ source-relay channel and $\mathbf{n}_{R,1}$ is the $N \times 1$ noise vector at the relay node. The relay processes \mathbf{y}_R through multiplication by a $N \times N$ matrix \mathbf{G} , retransmitting the symbol $\mathbf{x}_R = \mathbf{G}\mathbf{y}_R$ in the second time-slot. At the same time, the destination transmits a new symbol \mathbf{x}_2 , independent from \mathbf{x}_1 . Therefore, during the second time-slot the destination receives a symbol $\mathbf{y}_2 = \mathbf{H}_0\mathbf{x}_2 + \mathbf{H}_2\mathbf{x}_R + \mathbf{n}_{D,2}$, where \mathbf{H}_2 is the relay-destination channel and $\mathbf{n}_{D,2}$ is the $N \times 1$ noise vector at the destination. Finally, the overall input/output relation for this communication scheme is

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{H}_0 & \mathbf{0} \\ \mathbf{H}_2\mathbf{G}\mathbf{H}_1 & \mathbf{H}_0 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} + \mathbf{n}_{eq}, \quad (4.1)$$

where we have defined the equivalent noise vector as

$$\mathbf{n}_{eq} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_2\mathbf{G} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{n}_{D,1} \\ \mathbf{n}_{R,1} \\ \mathbf{n}_{D,2} \end{bmatrix}. \quad (4.2)$$

Since each noise vector is assumed to have a distribution $\mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$, the equivalent noise correlation matrix $\mathbf{C} = E[\mathbf{n}_{eq}\mathbf{n}_{eq}^H]$ is

$$\mathbf{C} = \begin{bmatrix} \sigma^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I} + \sigma^2\mathbf{H}_2\mathbf{G}\mathbf{G}^H\mathbf{H}_2^H \end{bmatrix} = \begin{bmatrix} \sigma^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2(\mathbf{G}) \end{bmatrix}. \quad (4.3)$$

Assuming for the sake of simplicity that \mathbf{x}_1 and \mathbf{x}_2 are independent, the input covariance matrix is block diagonal:

$$E[\mathbf{x}\mathbf{x}^H] = \mathbf{Q} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2 \end{bmatrix}. \quad (4.4)$$

Thanks to this assumption, this channel boils down to a two-users MIMO multiple

access channel. Thus, successive decoding of $(\mathbf{x}_1, \mathbf{x}_2)$ is a capacity-achieving decoding strategy and a full joint decoding is not needed.

4.2 Problem formulation

As in the previous section, we are interested in maximizing the mutual information between the source inputs $(\mathbf{x}_1, \mathbf{x}_2)$ and the destination outputs $(\mathbf{y}_1, \mathbf{y}_2)$, $I(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y}_1, \mathbf{y}_2)$, over the input covariance matrices, \mathbf{Q}_1 and \mathbf{Q}_2 , and the linear processing matrix at the relay, \mathbf{G} , under instantaneous power constraints for the source and relay input symbols. Notice that the ergodic achievable rate is $\frac{1}{2}E_{\mathcal{H}}[I(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y}_1, \mathbf{y}_2)]$. Thus we can formulate our optimization problem

$$\max_{\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{G}} \{I(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y}_1, \mathbf{y}_2)\} \quad (4.5a)$$

$$\text{s.t.} \begin{cases} \mathbf{Q}_i \succeq \mathbf{0} & i = 1, 2 \\ \text{tr}(\mathbf{Q}_i) = P_i & i = 1, 2 \\ \text{tr}(\mathbf{R}(\mathbf{Q}_1, \mathbf{G})) = P_R \end{cases} \quad (4.5b)$$

Given (4.4) and (4.1) the mutual information can be written as

$$I(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y}_1, \mathbf{y}_2) = C(\hat{\mathbf{H}}_1(\mathbf{G})\mathbf{Q}_1\hat{\mathbf{H}}_1^H(\mathbf{G})\mathbf{C}^{-1}(\mathbf{G}) + \hat{\mathbf{H}}_2\mathbf{Q}_2\hat{\mathbf{H}}_2^H\mathbf{C}^{-1}(\mathbf{G})) \quad (4.6)$$

where $\hat{\mathbf{H}}_1(\mathbf{G}) = \begin{bmatrix} \mathbf{H}_0 \\ \mathbf{H}_2\mathbf{G}\mathbf{H}_1 \end{bmatrix}$ and $\hat{\mathbf{H}}_2 = \begin{bmatrix} \mathbf{0} \\ \mathbf{H}_0 \end{bmatrix}$. In order to gain further information-theoretic insight on this optimization problem, we can use the chain rule [3] to expand the mutual information (4.6). Since $I(\mathbf{x}_2; \mathbf{y}_1) = 0$, we have

$$I(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y}_1, \mathbf{y}_2) = \underbrace{I(\mathbf{x}_1; \mathbf{y}_1)}_{I_1} + \underbrace{I(\mathbf{x}_1, \mathbf{x}_2; \mathbf{y}_2 | \mathbf{y}_1)}_{I_R} + \underbrace{I(\mathbf{x}_2; \mathbf{y}_2 | \mathbf{x}_1)}_{I_2}. \quad (4.7)$$

The total mutual information is the sum of three terms: the first term I_1 relates to the source transmission in the first time-slot, the second term I_R accounts for the signal re-transmitted by the relay in the second time-slot, while the third term I_2 depends on the source transmission in the second time-slot.

Recalling (4.1), the terms in (4.7) can be evaluated as follows¹

$$\begin{aligned} I_1 &= C(\mathbf{H}_0 \mathbf{Q}_1 \mathbf{H}_0^H) \\ I_R &= C\left(\mathbf{H}_2 \mathbf{G} \tilde{\mathbf{A}}(\mathbf{Q}_1) \tilde{\mathbf{A}}^H(\mathbf{Q}_1) \mathbf{G}^H \mathbf{H}_2^H \mathbf{B}^{-1}(\mathbf{G}, \mathbf{Q}_2)\right) \\ I_2 &= C(\mathbf{H}_0 \mathbf{Q}_2 \mathbf{H}_0^H \mathbf{C}_2^{-1}(\mathbf{G})), \end{aligned} \quad (4.8)$$

where the $N \times N$ matrix $\tilde{\mathbf{A}}(\mathbf{Q}_1)$ is

$$\tilde{\mathbf{A}}(\mathbf{Q}_1) = \mathbf{H}_1(\mathbf{Q}_1^{-1} + \mathbf{H}_0^H \mathbf{H}_0)^{-\frac{1}{2}}, \quad (4.9)$$

and the $N \times N$ matrix $\mathbf{B}(\mathbf{G}, \mathbf{Q}_2)$ accounts for all the interference and noise on the channel between the relay and the destination in the second time-slot:

$$\mathbf{B} = \sigma^2 \mathbf{I} + \sigma^2 \mathbf{H}_2 \mathbf{G} \mathbf{G}^H \mathbf{H}_2^H + \mathbf{H}_0 \mathbf{Q}_2 \mathbf{H}_0^H. \quad (4.10)$$

4.3 An iterative solution

The optimization problem (4.5) is not convex. However, if we fix either $(\mathbf{Q}_1, \mathbf{Q}_2)$ or \mathbf{G} , the resulting problem is convex in the remaining variable. Therefore, similarly to the previous section, we can devise an iterative procedure that alternates between the optimization over \mathbf{G} fixed $(\mathbf{Q}_1, \mathbf{Q}_2)$ and the optimization over $(\mathbf{Q}_1, \mathbf{Q}_2)$ fixed \mathbf{G} . Again, absolute convergence for this algorithm cannot be proved, since the power constraint on the relay input symbol depends upon *both* \mathbf{Q}_1 and \mathbf{G} . Nevertheless, if the problem is well-conditioned the algorithm has shown in practice rapid convergence (around five iterations) to a unique solution.

¹Notably, we could have obtained (4.8) also using the properties of the determinant of block matrices. Anyway, we preferred to present this result according to a more insightful information-theoretic approach.

4.3.1 The optimization of the source input covariance matrix

Let us first fix \mathbf{G} . The resulting optimization problem over $(\mathbf{Q}_1, \mathbf{Q}_2)$ is convex. As we did in the previous section, for the sake of our algorithm, the power constraint on the relay input symbol can be ignored. Resorting to the mutual information expression found in (4.6), our statement of the problem is

$$\max_{\mathbf{Q}_1, \mathbf{Q}_2} C(\hat{\mathbf{H}}_1 \mathbf{Q}_1 \hat{\mathbf{H}}_1^H \mathbf{C}^{-1} + \hat{\mathbf{H}}_2 \mathbf{Q}_2 \hat{\mathbf{H}}_2^H \mathbf{C}^{-1}) \quad (4.11a)$$

$$\text{s.t.} \begin{cases} \mathbf{Q}_i \succeq \mathbf{0} & i = 1, 2 \\ \text{tr}(\mathbf{Q}_i) = P_i & i = 1, 2 \end{cases} \quad (4.11b)$$

This problem is identical to finding the optimal input covariance matrices for a two-users MIMO-MAC with channels $\hat{\mathbf{H}}_1$ and $\hat{\mathbf{H}}_2$ and noise covariance matrix \mathbf{C} . The solution to this problem is the iterative-waterfilling algorithm proposed in [32].

4.3.2 The optimization of the linear processing at the relay

Let us now fix $\mathbf{Q}_1, \mathbf{Q}_2$. In this case, it is better to use the mutual information expansion found in (4.8), since it fully reveals the role \mathbf{G} plays in this scheme. Obviously, \mathbf{G} affects only the second time-slot terms. Focusing on these last two terms, it appears to be very hard to find an analytical expression for a matrix \mathbf{G} capable of maximizing *jointly* the relay mutual information term I_R and the source mutual information term I_2 . However, we observe that, as further detailed in Sec. 4.3.4, under appropriate conditions (namely a sufficiently good channel between source and relay) the term I_2 can be neglected without relevant performance loss. Therefore, in the following we focus on maximizing I_R . As a result, our assessment of the problem for the optimization of \mathbf{G} reads

$$\begin{aligned} & \max_{\mathbf{G}} \left\{ C(\mathbf{H}_2 \mathbf{G} \tilde{\mathbf{A}} \tilde{\mathbf{A}}^H \mathbf{G}^H \mathbf{H}_2^H \mathbf{B}^{-1}) \right\} \quad (4.12) \\ \text{s.t.} \quad & \text{tr}(\mathbf{G}(\mathbf{A}\mathbf{A}^H + \sigma^2 \mathbf{I})\mathbf{G}^H) = P_R, \end{aligned}$$

where $\mathbf{A} = \mathbf{H}_1 \mathbf{Q}_1^{\frac{1}{2}}$. We can further simplify the objective function extracting the interference term from the expression of \mathbf{B} (4.10)

$$\mathbf{B} = \tilde{\mathbf{Q}}_2^{\frac{1}{2}} (\sigma^2 \mathbf{I} + \sigma^2 \tilde{\mathbf{Q}}_2^{-\frac{1}{2}} \mathbf{H}_2 \mathbf{G} \mathbf{G}^H \mathbf{H}_2^H \tilde{\mathbf{Q}}_2^{-\frac{H}{2}}) \tilde{\mathbf{Q}}_2^{\frac{H}{2}},$$

where $\tilde{\mathbf{Q}}_2 = \mathbf{I} + \frac{1}{\sigma^2} \mathbf{H}_0 \mathbf{Q}_2 \mathbf{H}_0^H$. Defining $\tilde{\mathbf{H}}_2 = \tilde{\mathbf{Q}}_2^{-\frac{1}{2}} \mathbf{H}_2$ and using the commutative property of the determinant $\det(\mathbf{I} + \mathbf{A}\mathbf{B}) = \det(\mathbf{I} + \mathbf{B}\mathbf{A})$, we can re-write (4.12) as

$$\max_{\mathbf{G}} \left\{ C(\tilde{\mathbf{H}}_2 \mathbf{G} \tilde{\mathbf{A}} \tilde{\mathbf{A}}^H \mathbf{G}^H \tilde{\mathbf{H}}_2^H (\sigma^2 \mathbf{I} + \sigma^2 \tilde{\mathbf{H}}_2 \mathbf{G} \mathbf{G}^H \tilde{\mathbf{H}}_2^H)^{-1}) \right\}$$

The objective function is now the same as in (2.6), so we can use the same arguments. In particular, we choose $\mathbf{G} = \tilde{\mathbf{V}}_2 \tilde{\mathbf{D}}_g \tilde{\mathbf{U}}_A^H$, where $\tilde{\mathbf{D}}_g$ is diagonal, $\tilde{\mathbf{V}}_2$ is the matrix of the right eigenvectors of $\tilde{\mathbf{H}}_2$ and $\tilde{\mathbf{U}}_A$ is the matrix of the left eigenvectors of $\tilde{\mathbf{A}}$. We can now write the diagonalized optimization problem in standard form (as defined in [28])

$$\begin{aligned} \min_{\{\tilde{g}_1, \dots, \tilde{g}_N\}} & \left\{ - \sum_{r=1}^N \log \left(1 + \frac{\tilde{\lambda}_{A,r}^2 \tilde{\lambda}_{2,r}^2 |\tilde{g}_r|^2}{\sigma^2 + \sigma^2 \tilde{\lambda}_{2,r}^2 |\tilde{g}_r|^2} \right) \right\} \\ \text{s.t.} & \sum_{r=1}^N (a_{rr} + \sigma^2) |\tilde{g}_r|^2 = P_R \end{aligned} \quad (4.13)$$

where $\tilde{\lambda}_{A,r}$ and $\tilde{\lambda}_{2,r}$ are the r -th singular value of $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{H}}_2$, g_r is the r -th diagonal element of matrix $\tilde{\mathbf{G}}$, and a_{rr} is the nonnegative r -th diagonal element of the semi-definite positive matrix $\tilde{\mathbf{U}}_A^H \mathbf{A} \mathbf{A}^H \tilde{\mathbf{U}}_A$. The solution found solving the KKT condition is

$$|\tilde{g}_r|^2 = \frac{1}{\tilde{\lambda}_{A,r}^2 + \sigma^2} \left[\tilde{f}(\mu; \tilde{\eta}_r) - \frac{\sigma^2}{\tilde{\lambda}_{2,r}^2} \right]^+ \quad (4.14)$$

$$\tilde{f}(\mu; \tilde{\eta}_r) = \sqrt{\left(\frac{\tilde{\eta}_r}{2}\right)^2 + \frac{\tilde{\lambda}_{A,r}^2 + \sigma^2}{a_{rr} + \sigma^2} \tilde{\eta}_r \mu^2} - \frac{\tilde{\eta}_r}{2}, \quad (4.15)$$

where $\tilde{\eta}_r = \frac{\tilde{\lambda}_{A,r}^2}{\tilde{\lambda}_{2,r}^2}$. The value of μ is chosen as to satisfy the power constraint on the relay input symbol $\sum_{r=1}^N (a_{rr} + \sigma^2) |g_r|^2 = P_R$.

4.3.3 Algorithm definition

Following from this results, we can finally detail our algorithm:

```

initialize  $\mathbf{Q}_1 = \mathbf{Q}_2 = \mathbf{0}$  ,  $\mathbf{G} = \mathbf{I}$ ;
repeat
(a) solve the optimization problem (4.11) for  $(\mathbf{Q}_1, \mathbf{Q}_2)$  keeping  $\mathbf{G}$  fixed;
(b) solve the problem of optimizing  $I_R$  over  $\mathbf{G}$  keeping  $(\mathbf{Q}_1, \mathbf{Q}_2)$  fixed;
until the rate (4.6) converges.

```

We choose to initialize $\mathbf{G} = \mathbf{I}$, because otherwise there would be only a direct link between the source and the destination at the starting point of the optimization. Notice that the last iteration of the algorithm has to be the optimization of I_R over \mathbf{G} , since this guarantees the enforcement of all the constraints in the original problem (4.5). As we said before, absolute convergence to an optimal solution is not guaranteed for this algorithm. However, if the problem is well-conditioned, it has shown to rapidly converge² to a unique sub-optimal solution even from a randomly chosen starting point. This algorithm is efficient in that it decomposes the general non-convex multi-user problem into a sequence of convex single-user problems, each of which is much simpler to solve.

4.3.4 Simulation results

In this section, we assume that the relay is located on a line between the source and the destination, at a normalized distance $d \in [0, 1]$ from the source and $(1-d)$ from the destination. It follows that, assuming a path-loss exponent of 4, the channel matrix \mathbf{H}_i has entries distributed as $\mathcal{CN}(0, d_i^{-4})$, where $d_1 = d$, $d_2 = (1-d)$ and $d_0 = 1$. As far as the power constraints are concerned, we fix $P_1 = 1$ and $P_2 = P_R = \frac{1}{2}$, so as to satisfy a per-slot sum-power constraint $P_2 + P_R = 1$ and obtain fair performance comparison.

²around 5-6 iterations with a tolerance of 10^{-2} on the rate

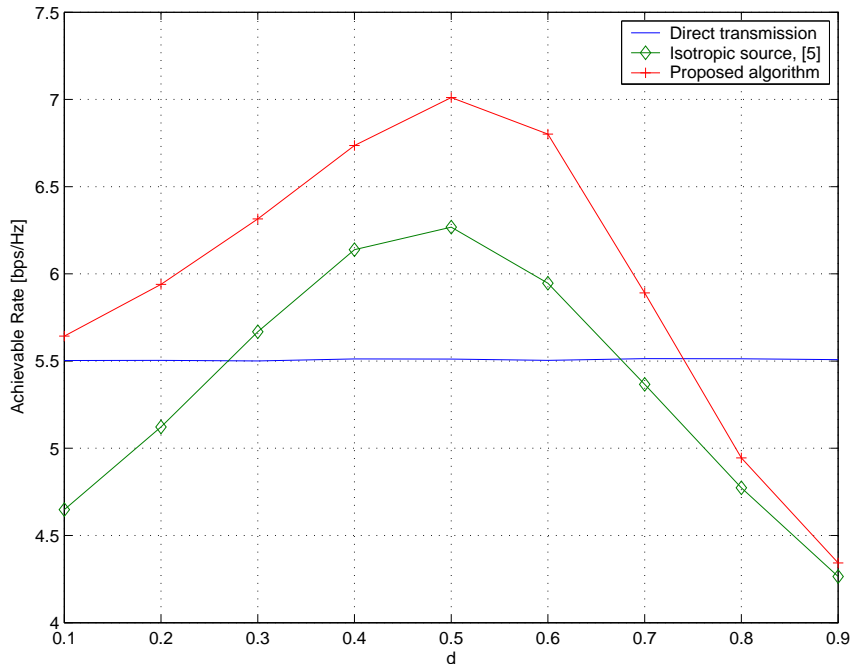


Figure 4.2: Achievable rates of different communication schemes for the cooperative system with $\frac{1}{\sigma^2} = 5dB$, $N = 3$.

The achievable rates of different algorithms for the cooperative scenario are depicted in fig. 4.2 versus the distance d for $\frac{1}{\sigma^2} = 5dB$ and a number of antennas at each node $N = 3$. In particular, both the technique presented in [5] (that assumes an isotropic covariance matrix $\mathbf{Q} = N^{-1}\mathbf{I}$) and the proposed method are compared to the reference performance of direct transmission (between source and destination). In this case, the proposed algorithm outperforms both direct transmission (by up to 1.5 bps/Hz) and the scheme of [5] (by up to 0.8 bps/Hz) for almost every value of d .

It seems now reasonable to ask whether the performance boost of our scheme is due to the protocol used or to the iterative optimization of the input covariance matrix and relay linear processing matrix. In fig. 4.3 we have plotted the achievable rates for an iterative algorithm of the same fashion of the one we proposed, but applied to the protocol used in [5]. It is seen that there is no gain in performing an iterative

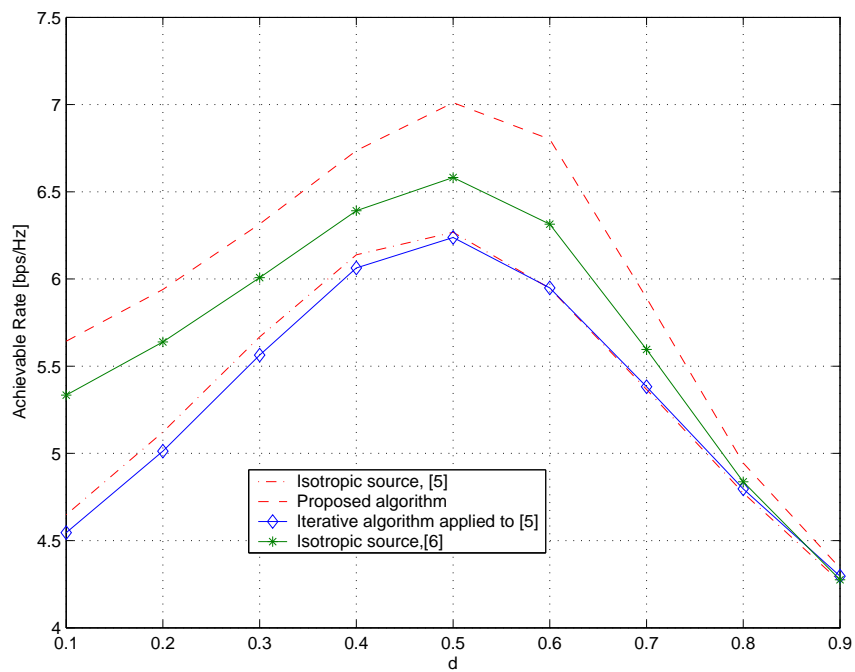


Figure 4.3: Achievable rates of different communication schemes for the cooperative system with $\frac{1}{\sigma^2} = 5dB$, $N = 3$.

optimization for this particular protocol. This result is in a way similar to the multi-hop case seen in Sec. 2.3.4, but here the iterative procedure does not attain any practical gain for $d > 0.5$. Our intuition is that the iterative algorithm suffers from the non-square channel structure of the protocol employed in [5]. The other curve is the achievable rate optimizing only the matrix \mathbf{G} (as detailed in Sec. 4.3.2) employing the same protocol described in Sec. 4.1. While this strategy clearly outperforms the scheme of [5], its rate is still well under the rate of the algorithm proposed in this Section. In conclusion, it seems that the reason of the performance boost of our scheme is due to *both* the protocol employed and the iterative optimization.

Finally, we would like to comment on the choice of neglecting the term I_2 while optimizing for \mathbf{G} in Sec. 4.3. Fig. 4.4 shows I_2 versus d for the proposed scheme, along with the upper bound $I_2 \leq C(\frac{1}{\sigma^2}\mathbf{H}_0\mathbf{Q}_2\mathbf{H}_0^H) = I_{2ub}$. It is seen that for $d \leq 0.4$ (i.e., for a sufficiently good channel between source and relay), there is no optimality loss due to I_2 . However, for $d \geq 0.4$ it is envisaged that a *joint* optimization of I_2 and I_R over \mathbf{G} might bring performance benefits.

4.4 Low-complexity iterative solution

Similarly to Sec. 2.4 in the multi-hop case, in this Section we present a low-complexity iterative algorithm, that finds an approximation of the exact solution found through the iterative algorithm of Sec. 4.3. Building on the results of Sec. 2.4, both the optimizations performed in each step of the iterative algorithm can be approximated using a constant-power water-filling allocation. In fact, in each step of the iterative-waterfilling procedure [32] required to solve (4.11) two simple single-user water-filling are performed, and so the constant-power algorithm of [29] can be used. As far as the optimization of I_R is concerned, we can use the same arguments as in Sec. 2.4.1. The low-complexity iterative algorithm can be then detailed as

```

initialize  $\mathbf{Q}_1 = \mathbf{Q}_2 = \mathbf{0}$  ,  $\mathbf{G} = \mathbf{I}$ ;
repeat
(a) use the constant-power allocation algorithm in [29] to find a

```

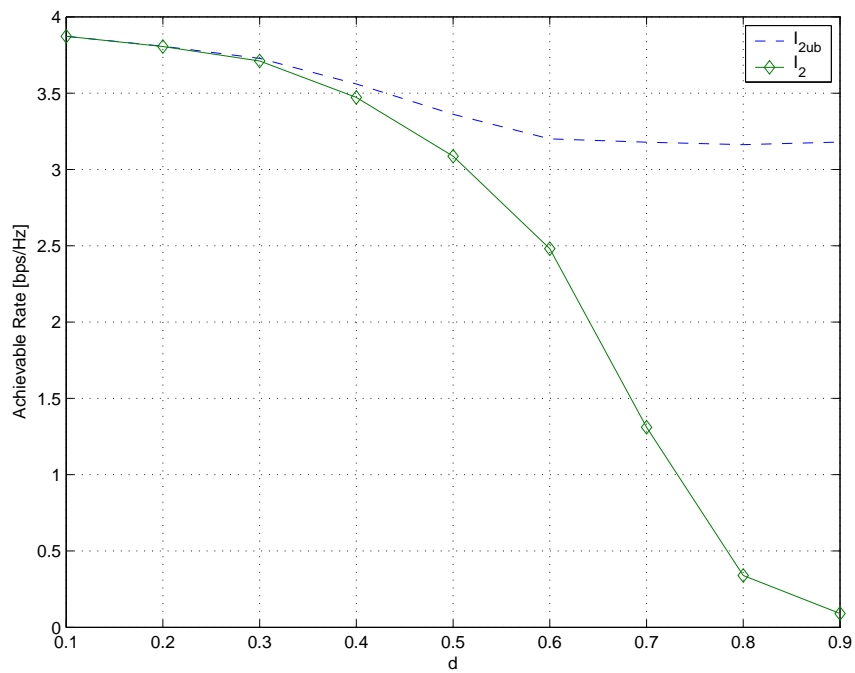


Figure 4.4: Rate component I_2 and upper bound I_{2ub} as a function of d for $\frac{1}{\sigma^2} = 5dB$, $N = 3$.

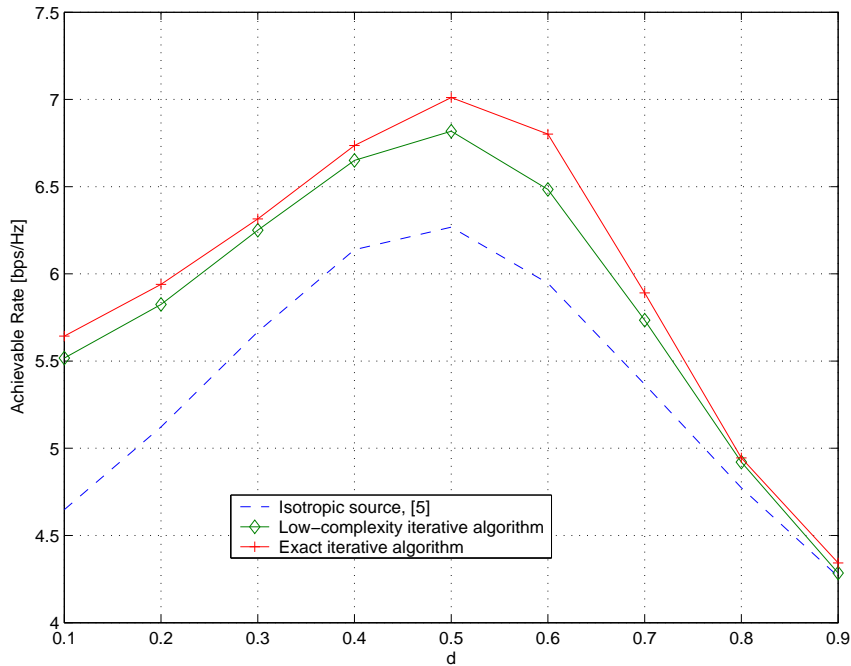


Figure 4.5: Achievable rate of the low-complexity iterative algorithm compared with the exact iterative algorithm presented in Sec. 4.3 and the communication scheme of [5], $N = 3$, $\frac{1}{\sigma^2} = 5dB$.

sub-optimal solution to (4.11) for $(\mathbf{Q}_1, \mathbf{Q}_2)$ keeping \mathbf{G} fixed;
 (b) use the constant-power allocation algorithm presented in Sec. 2.4.2 to find a sub-optimal solution to the problem of optimizing I_R over \mathbf{G} keeping $(\mathbf{Q}_1, \mathbf{Q}_2)$ fixed;
 until the rate (4.6) converges.

It is not possible to prove absolute convergence for this algorithm. Nevertheless, as the algorithm presented in Sec. 4.3, if the problem is well-conditioned, it has shown to rapidly converge³ to a unique sub-optimal solution even from a randomly chosen starting point.

The ergodic achievable rate of the low-complexity iterative algorithm is shown in fig. 4.5 and compared with the achievable rates of the exact algorithm presented in

³around 5-6 iterations with a tolerance of 10^{-2} on the rate.

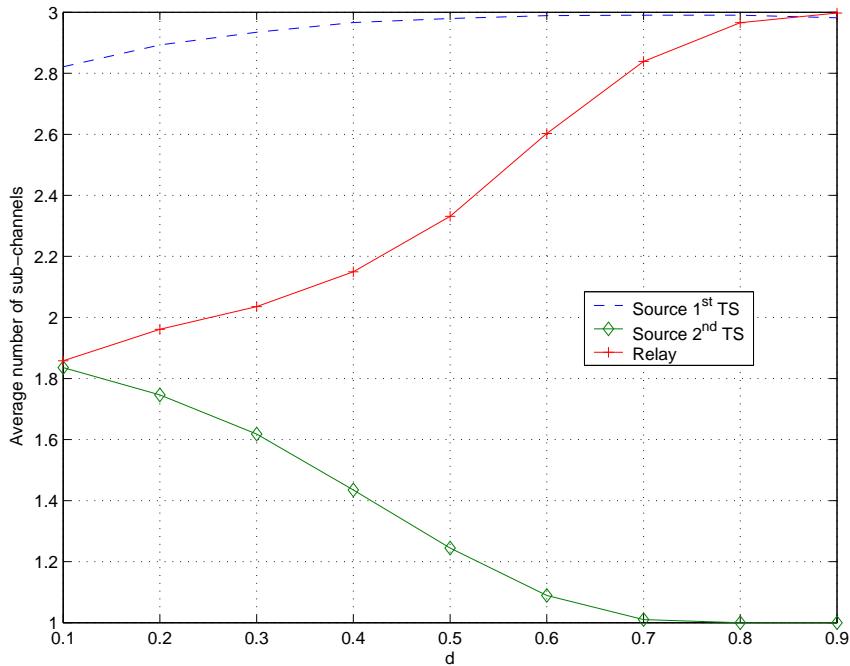


Figure 4.6: Number of sub-channels used by the relay and source, according to the low-complexity iterative algorithm, $N = 3$, $\frac{1}{\sigma^2} = 5dB$.

Sec. 4.3 and the communication scheme of [5], for $N = 3$, $\frac{1}{\sigma^2} = 5dB$ and $P_1 = 1$, $P_2 = 1/2$ and $P_R = 1/2$. It is seen that the rate of the low-complexity algorithm is close to the rate of the exact algorithm for a wide range of values of d and outperforms the rate of [5] for all values of d .

In fig. 4.6 the number of spatial sub-channels used by the source and the relay is shown on average with respect to the channels realizations. It is seen that the number of channels used by the source in the second time slot decreases as the interference from the relay grows. Differently from the multi-hop case, the number of channels used by the source in the first time-slot is always close to the maximum of 3. Our intuition is that this is due to the use of the collaborative scheme.

One-iteration performance

As we did for the multihop case, it is of great interest for the practical implementation of the low-complexity algorithm previously presented to analyze its performance after

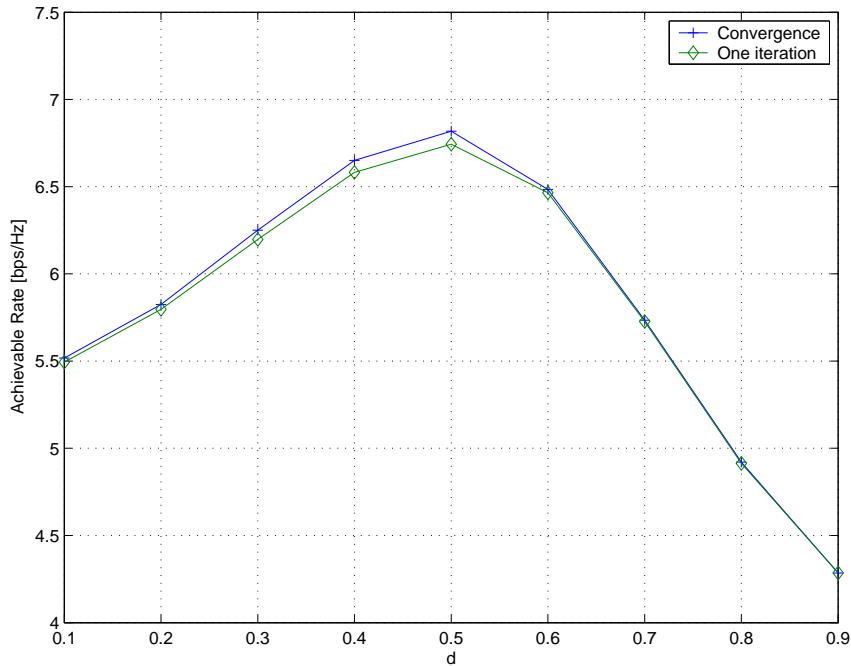


Figure 4.7: Achievable rate of the low-complexity iterative algorithm at convergence and after one iteration, $N = 3$, $\frac{1}{\sigma^2} = 5dB$.

just one step of the iterative procedure. Of course, also one single step of the iterative-waterfilling procedure has to be performed. In this way, the computation of only four SVD will be required: one to find the signaling sub-space for \mathbf{Q}_1 , one to find the signaling sub-space for \mathbf{Q}_2 , and two to find the sub-spaces used by the relay gain matrix \mathbf{G} . The average achievable rate after one iteration is shown in fig. 4.7 versus the achievable rate at convergence for $N = 3$, $\frac{1}{\sigma^2} = 5dB$. For $d < 0.6$ only a fraction of bit/s/Hz is lost, while for $d > 0.6$ we have immediate convergence, as in the multi-hop case. Further, in fig. 4.8 it is seen that, after just one iteration, the low-complexity iterative algorithm performs better than the scheme in [5] for all the values of d and it is also rather close to the rate of the iterative algorithm of Sec. 4.3.

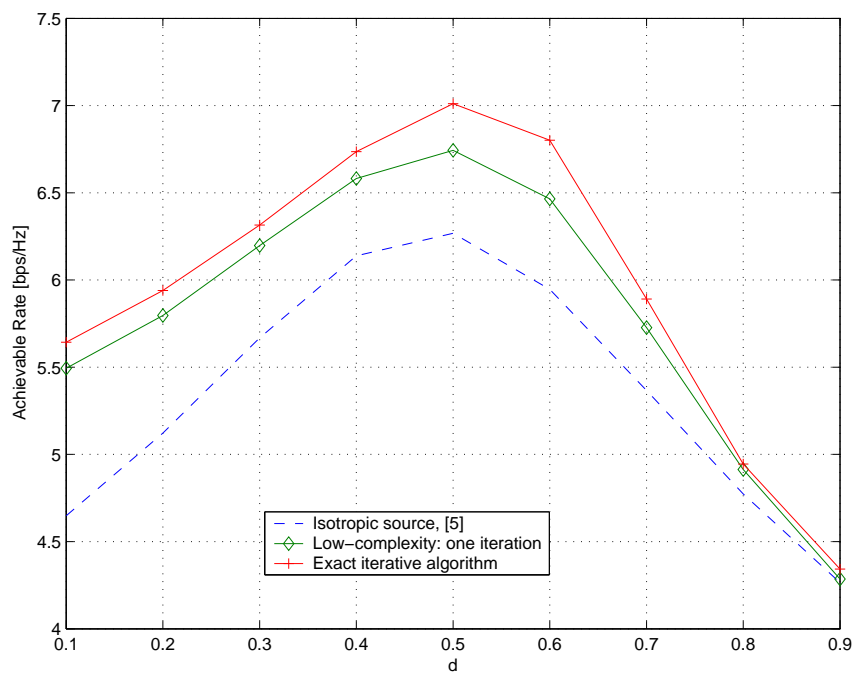


Figure 4.8: Achievable rate of the low-complexity iterative algorithm after one iteration compared to the scheme in [5] and to the exact iterative algorithm, $N = 3$, $\frac{1}{\sigma^2} = 5dB$.

Chapter 5

Conclusions

In this work, the potential benefits of deploying multiple antennas at the nodes of an AF relay channel have been investigated. In particular, two types of MIMO Amplify-and-Forward relay systems were considered: multi-hop and cooperative. For both cases, the problem of maximizing the achievable rate over the covariance matrices of the symbols transmitted by the source and relay linear processing matrix was formulated under the assumption of full channel state information at each node. A sub-optimal iterative algorithm has been proposed and proved by numerical simulations to outperform known schemes. A more accurate analysis of the relay optimization step enabled a low-complexity implementation of the overall iterative algorithm. The performance after just one-iteration of the low-complexity algorithm was shown to be not far from the rate achieved by the exact iterative algorithm.

Also, optimal power allocation and linear processing have been investigated for an amplify-and-forward fading relay channel with multiple antennas at the relay node and under an instantaneous sum-power constraint. In this scenario, the optimal linear processing at the relay node is the outer product of the beamformers for the source-relay and relay-destination channels. Moreover, the optimal transmission scheme is either direct transmission (the relay remains silent in the second time slot), or the scheme proposed in [22] (the source remains silent in the second time-slot).

Appendix A

Convex Optimization Fundamentals

Recently, convex optimization has become a fundamental tool to solve constrained optimization problems in a variety of fields ranging from automatic control to communications. In this Chapter, basic convex optimization tools will be introduced. The main reference for the results presented in the following is [28], which is the most modern and complete reference on the topic.

A.1 General optimization problems and the duality gap

A general $\mathbb{R}^N \rightarrow \mathbb{R}$ optimization problem can be stated in the following form

$$\min_{\mathbf{x}} f_0(\mathbf{x}) \tag{A.1}$$

$$\text{s.t.} \begin{cases} f_i(\mathbf{x}) \leq 0 \\ h_j(\mathbf{x}) = 0 \end{cases}, \tag{A.2}$$

where $f_0(\mathbf{x})$ is the objective function; $f_i(\mathbf{x})$, $i = 1, \dots, n$ are the inequality constraints functions, and $h_j(\mathbf{x})$, $j = 1, \dots, m$ are the equality constraints functions. The Lagrangian of the optimization problem is defined as

$$L(\mathbf{x}, \boldsymbol{\gamma}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_{i=1}^n \gamma_i f_i(\mathbf{x}) + \sum_{j=1}^m \nu_j h_j(\mathbf{x}), \quad (\text{A.3})$$

where γ_i are positive constants. The Lagrangian dual function is defined as [28]

$$g(\boldsymbol{\gamma}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\gamma}, \boldsymbol{\nu}), \quad (\text{A.4})$$

and depends only on the dual variables $(\boldsymbol{\gamma}, \boldsymbol{\nu})$. It is easy to see [28] that $g(\boldsymbol{\gamma}, \boldsymbol{\nu})$ is a *lower bound* on the optimal value of $f_0(\mathbf{x})$, that is

$$g(\boldsymbol{\gamma}, \boldsymbol{\nu}) \leq \min_{\mathbf{x}} f_0(\mathbf{x}). \quad (\text{A.5})$$

The difference between the primal objective and the Lagrange dual is called the *duality gap*

$$\Gamma = f_0(\mathbf{x}^*) - g(\boldsymbol{\gamma}, \boldsymbol{\nu}). \quad (\text{A.6})$$

Given a sub-optimal solution \mathbf{x}^* to the optimization problem (A.2), the duality gap is therefore a *theoretical* upper bound to the error

$$\epsilon = |f_0(\mathbf{x}^*) - \min_{\mathbf{x}} f_0(\mathbf{x})| \leq |f_0(\mathbf{x}^*) - g(\boldsymbol{\gamma}, \boldsymbol{\nu})|. \quad (\text{A.7})$$

A.2 Convex functions

A function $f : \mathbb{R}^N \rightarrow \mathbb{R}$ is convex if the domain of f , $\mathbf{dom} f$, is a convex set and if for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom} f$, $\theta \in [0, 1]$

$$f(\theta \mathbf{x} + (1 - \theta) \mathbf{y}) \leq \theta f(\mathbf{x}) + (1 - \theta) f(\mathbf{y}). \quad (\text{A.8})$$

Geometrically, this inequality means that the line segment between $(\mathbf{x}; f(\mathbf{x}))$ and $(\mathbf{y}; f(\mathbf{y}))$, which is the chord from \mathbf{x} to \mathbf{y} , lies above the graph of f . A function is convex if and only if it is convex when restricted to any line that intersects its domain. In other words f is convex if and only if for all $\mathbf{x} \in \mathbf{dom} f$ and all \mathbf{v} , the function $g(t) = f(\mathbf{x} + t\mathbf{v})$ is convex (on its domain, $\{t | \mathbf{x} + t\mathbf{v} \in \mathbf{dom} f\}$). This property is very useful, since it allows us to check whether a function is convex by restricting it to a line.

A.3 Convex optimization problems

The problem (A.2) is convex if $f_i(\mathbf{x})$, $i = 0, \dots, n$ are convex and $h_j(\mathbf{x})$, $i = 1, \dots, m$ are affine. In this case, if Slater's *constraint qualifications* hold (which is almost always the case), the duality gap at the optimum is zero. So, we can use the duality gap (A.6) to investigate the optimality of a general sub-optimal solution \mathbf{x}^* .

Various flavors of optimality conditions have been found for convex optimization problems. In particular, for any convex optimization problem with differentiable objective and constraint functions the Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient for the points \mathbf{x}^* and $(\boldsymbol{\gamma}^*, \boldsymbol{\nu}^*)$ to be primal and dual optimal

$$f_i(\mathbf{x}^*) \leq 0 \quad i = 1, \dots, n \quad (\text{A.9})$$

$$h_j(\mathbf{x}^*) = 0 \quad i = 1, \dots, m \quad (\text{A.10})$$

$$\gamma_i^* > 0 \quad i = 1, \dots, n \quad (\text{A.11})$$

$$\gamma_i^* f_i(\mathbf{x}^*) = 0 \quad i = 1, \dots, n \quad (\text{A.12})$$

$$\nabla f_0(\mathbf{x}^*) + \sum_{i=1}^n \gamma_i^* \nabla f_i(\mathbf{x}^*) + \sum_{j=1}^m \nu_j^* \nabla h_j(\mathbf{x}^*) = 0 \quad (\text{A.13})$$

The KKT conditions play an important role in optimization. In a few special cases it is possible to solve the KKT conditions (and therefore, the optimization problem) analytically. More generally, many algorithms for convex optimization are conceived as, or can be interpreted as, methods for solving the KKT conditions.

Bibliography

- [1] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, pp.585-595, Nov. 1999.
- [2] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Commun.: Kluwer Academic Press*, no. 6, pp. 311335, 1998.
- [3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Interscience, 1991.
- [4] A. Goldsmith, P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Inf. Theory*, vol. 43, pp. 1986-1992, Nov. 1997.
- [5] O. Munoz, J. Vidal and A. Agustin, "Non-regenerative MIMO relaying with channel state information," *Proceedings ICASSP '05*, vol. 3, pp. 361-364, March 2005.
- [6] R. U. Nabar, H. Bölcskei, and F. W. Kneubuhler, "Fading relay channels: Performance limits and space-time signal design," *IEEE Jour. on Selected Areas in Communications*, vol. 22, Issue 6, pp. 1099-1109, June 2004.
- [7] A. Goldsmith and P. Varaiya, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. Communications*, pp. 1986-1992, Nov. 1997.
- [8] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE Jour. on Selected Areas Communication* vol. 16pp. 1451-1458, Oct. 1998.

- [9] L. Zheng and D. Tse, "Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 49, pp. 1073-1096, May 2003.
- [10] C. E. Shannon, W. Weaver, *A mathematical theory of communication*, University of Illinois Press, 1963.
- [11] H. Weingarten, Y. Steinberg, S. Shamai, "The capacity region of the Gaussian MIMO broadcast channel," in *Proc ISIT 2004*.
- [12] H. Sato, "The capacity of the Gaussian interference channel under strong interference," *IEEE Trans. Inform. Theory*, vol. 27, pp. 786-788, Nov. 1981.
- [13] T. S. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Trans. Inform. Theory*, vol. 27, pp. 4960, Jan. 1981.
- [14] E. van der Meulen, "A survey of multi-way channels in information theory: 1961-1976," *IEEE Trans. Inf. Theory*, vol. 23, pp. 1-37, Jan. 1977.
- [15] T. M. Cover and A. A. El Gamal, "Capacity theorems for the relay channel," *IEEE Trans. Inf. Theory*, vol. 25, pp. 572-584, Sep. 1979.
- [16] A. Carleial, "Multiple-access channels with different generalized feedback signals," *IEEE Trans. Inf. Theory*, vol. 28, pp. 841-850, Nov. 1982.
- [17] F. Willems, E. van der Meulen "The discrete memoryless multiple-access channel with cribbing encoders," *IEEE Trans. Inf. Theory*, vol. 31, pp. 313-327, May 1985.
- [18] E. van der Meulen, "Three-terminal communication channels," *Adv. Appl. Pmb.*, vol. 3, pp.120-54, 1971.
- [19] M. Khojastepour, A. Sabharwal and B. Aazhang, "On the capacity of cheap relay networks," in *Proc. CISS 2003*.
- [20] M. Khojastepour, A. Sabharwal and B. Aazhang, "On capacity of Gaussian 'cheap' relay channel," in *Proc. GLOBECOM'03*.

- [21] Y. Liang and V. V. Veeravalli, "Gaussian Orthogonal Relay Channels: Optimal Resource Allocation and Capacity," *IEEE Trans. Inf. Theory*, vol. 51, pp. 3284-3289, Sept. 2005.
- [22] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behaviour," *IEEE Trans. Inf. Theory*, vol. 50, pp. 3062-3080, Dec. 2004.
- [23] Y. Yingwei, C. Xiaodong and G. B. Giannakis, "On energy efficiency and optimum resource allocation of relay transmissions in the low-power regime," *IEEE Trans. on Wireless Communications*, vol. 4, Issue 6, pp. 2917-2927, Nov. 2005.
- [24] L. Yingbin and V. V. Veeravalli, "Resource allocation for wireless relay channels," *38th Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1902-1906, 7-10 Nov. 2004 .
- [25] A. Host-Madsen and Z. Junshan, "Capacity bounds and power allocation for wireless relay channels", *IEEE Trans. Inf. Theory*, vol. 51, Issue 6, pp. 2020-2040, June 2005.
- [26] N. Ahmed, M. A. Khojastepour and B. Aazhang, "Outage minimization and optimal power control for the fading relay channel," *IEEE Information Theory Workshop*, pp. 458-462, 24-29 Oct. 2004. vol. 51, Issue 6, pp. 2020-2040, June 2005.
- [27] M. A. Khojastepour, A. Sabharwal and B. Aazhang, "Lower bounds on the capacity of Gaussian relay channel," *38th Annual Conf. on Info. Sci. and Sys. (CISS 2004)*, Princeton, NJ.
- [28] S. Boyd and L. Vanderberghe, *Convex Optimization*, Cambridge University Press, 2003.
- [29] W. Yu and J. M. Cioffi, "On constant power water-filling," in *Proc. IEEE ICC 2001*.

- [30] B. Wang, J. Zhang, A. Host-Madsen, “On the capacity of MIMO relay channels,” *IEEE Trans. Inf. Theory*, vol. 51, Issue 1, pp. 29 - 43, Jan. 2005.
- [31] D. Bertsekas, *Nonlinear Programming*, Athena Scientific, 1999.
- [32] W. Yu, W. Rhee, S. Boyd, and J. Cioffi. “Iterative water-filling for Gaussian vector multiple access channels,” *IEEE Trans. Inform. Theory*, vol. 50, pp. 145-152, Jan. 2004.
- [33] P. S. Chow, *Bandwidth optimized digital transmission techniques for spectrally shaped channels with impulse noise*, Ph.D. thesis, Stanford University, 1993.