CAPACITY LIMITS IN NON-UNIFORM CSI BROADCAST CHANNELS AND IN SPECTRUM SHARING NETWORKS

by

Yang Li

APPROVED BY SUPERVISORY COMMITTEE:

Dr. Aria Nosratinia, Chair

Dr. Naofal Al-Dhahir

Dr. Hlaing Minn

Dr. Won Namgoong

Dedicated to my parents,

and my wife, Xiaomin Chen.

CAPACITY LIMITS IN NON-UNIFORM CSI BROADCAST CHANNELS AND IN SPECTRUM SHARING NETWORKS

by

YANG LI, B.S., M.S.

DISSERTATION

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY IN ELECTRICAL ENGINEERING

THE UNIVERSITY OF TEXAS AT DALLAS

December 2012

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

ACKNOWLEDGMENTS

I want to thank my advisor, Dr. Aria Nosratinia, who worked closely with me throughout my four-year Ph.D. studies. In my first semester he told me: “Every great idea is simple.” These words motivated me to seek the most interesting ideas in the field of wireless communication, and to extract the essence of problems. I will never forget the weekends he worked with me in a quiet campus or in a coffee shop; I will never forget discussing ideas and research in his house over tea and cookies while he was holding his infant baby. I learned (and will keep learning) from his insights, passion for life and work, and persistence for perfection. I always remind myself of his words “everything we produce must have the highest quality.” He gave me all I could ever ask from an advisor.

I want to thank my committee members, Dr. Naofal Al-Dhahir, Dr. Hlaing Minn and Dr. Won Namgoong, for their support and advices. From them I learned different perspectives of viewing research problems.

I would also like to thank my colleagues at The University of Texas at Dallas, especially Ramy Tannious, Negar Bazargani, Mohamed AbolSoud, Ahmed Hesham, Thuy Nguyen and Baile Xie. I enjoyed and learned from the discussions and working with them.

Finally, I want to thank my wife and my parents for their wonderful support. My wife, Xiaomin Chen, has been a constant source of encouragement; I cannot say how much I appreciate her patience with my focus on research and, sometimes, my absent-mindedness. My parents have also been understanding and encouraged me to work hard, even though I was not able to visit them in China during the past four years. I am so blessed to have them in my life and without their support this dissertation would not have been possible.

November 2012

PREFACE

This dissertation was produced in accordance with guidelines which permit the inclusion as part of the dissertation the text of an original paper or papers submitted for publication. The dissertation must still conform to all other requirements explained in the "Guide for the Preparation of Master's Theses and Doctoral Dissertations at The University of Texas at Dallas." It must include a comprehensive abstract, a full introduction and literature review, and a final overall conclusion. Additional material (procedural and design data as well as descriptions of equipment) must be provided in sufficient detail to allow a clear and precise judgment to be made of the importance and originality of the research reported.

It is acceptable for this dissertation to include as chapters authentic copies of papers already published, provided these meet type size, margin, and legibility requirements. In such cases, connecting texts which provide logical bridges between different manuscripts are mandatory. Where the student is not the sole author of a manuscript, the student is required to make an explicit statement in the introductory material to that manuscript describing the student's contribution to the work and acknowledging the contribution of the other author(s). The signatures of the Supervising Committee which precede all other material in the dissertation attest to the accuracy of this statement.

CAPACITY LIMITS IN NON-UNIFORM CSI BROADCAST CHANNELS AND IN SPECTRUM SHARING NETWORKS

Publication No. ___________________

Yang Li, B.S., M.S.
The University of Texas at Dallas, 2012

Supervising Professor: Dr. Aria Nosratinia

This dissertation has two parts: the first part studies multi-antenna broadcast channels with nodes of varying mobility, and the second part studies capacity limits of spectrum-sharing networks.

In the multi-antenna broadcast channel without transmit-side channel state information (CSIT), it has been known that when all receivers have channel state information (CSIR), the degrees of freedom (DoF) cannot be improved beyond what is available via TDMA. The same is true if none of the receivers possess CSIR. This dissertation shows that an entirely new scenario emerges when receivers have unequal CSIR: orthogonal transmission is no longer DoF-optimal. In particular, when one receiver has CSIR and the other does not, two product superposition methods based on Grassmannian signaling are proposed and analyzed, and are shown to attain the optimal degrees of freedom for a wide set of antenna configurations and channel coherence times. Furthermore, the product superposition is extended to the domain of coherent signaling with pilots, the advantages of product superposition are demonstrated in low-SNR as well as high-SNR, and DoF optimality is established in a wider set of receiver antenna configurations. Two classes of decoders, with and without interference cancellation, are studied, and the effect of power allocation and partial CSI at the base station are investigated.

The second part of this dissertation investigates capacity limits of spectrum-sharing networks. Unlike point-to-point cognitive radio, where the constraint imposed by the primary rigidly curbs the secondary throughput, multiple secondary users have the potential to more efficiently harvest the spectrum and share it among themselves. Efficient methods are proposed and analyzed to exploit multiuser diversity in cognitive broadcast channels, cognitive multiple access channel (MAC) and cognitive relay channels. The optimal growth rate of the capacity of these channels is established, and the tradeoff between scaling the secondary throughput and reducing interference on the primary is highlighted and characterized.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Background

Wireless communication has attracted significant interests in the past decades. To sustain ever increasing demand for mobile data, various advanced wireless technologies have been developed to maximize spectrum efficiency. Among these technologies, multiple-input-multiple-output (MIMO) has shown great potential to improve data rate and are widely used by standards, e.g., Long Term Evolution (LTE) and IEEE 802.22.

MIMO systems use multiple antennas at transmitter and receiver and may send multiple data streams to one or several receivers at the same time and frequency, thus the overall spectral efficiency can be greatly increased. To harvest the gain of MIMO, the receiver(s) must reliably estimate the channel to the transmitter, and, if necessary, feed back the channel state information (CSI) to the transmitter in a timely fashion. It is widely understood that CSI acquisition is costly in resources and sometimes may be challenging or even infeasible, and therefore is one key of issues that limit the gain promised by MIMO.

Cognitive radio (CR) aims to improve spectral efficiency from a different perspective. It is known that the spectrum assigned to licensed (primary) users is severely under-utilized [1]. The utilization of spectrum can be increased by allowing cognitive radio (secondary) users to access the same spectrum as primary users, as long as performance degradation of the primary users remains acceptable. In general the secondary users can only use the spectrum when the spectrum demand of primary users is not heavy or the interference between the primary and secondary users is small due to, e.g., channel fading or spatial separation. It is challenging for the secondary users to fulfill constraints imposed by the primary system while maximizing their own data rate.

1.2 Motivations and Objectives

This dissertation studies the MIMO broadcast channels and cognitive radio networks.

1.2.1 MIMO Broadcast Channels

In the multi-antenna broadcast channel without transmit-side channel state information (CSIT), it has been known that when all receivers have channel state information (CSIR), the degrees of freedom (DoF) cannot be improved beyond what is available via TDMA. The same is true if none of the receivers possess CSIR. This dissertation shows that an entirely new scenario emerges when receivers have unequal CSIR. In particular, orthogonal transmission is no longer DoF-optimal when one receiver has CSIR and the other does not. A multiplicative superposition is proposed for this scenario and shown to attain the optimal degrees of freedom under a wide set of antenna configurations and coherence lengths. The product superposition is extended to the domain of coherent signaling with pilots, the advantages of product superposition are demonstrated in low-SNR as well as high-SNR.

1.2.2 Cognitive Radio Networks

Unlike point-to-point cognitive radio, where the constraint imposed by the primary rigidly curbs the secondary throughput, multiple secondary users have the potential to more efficiently harvest the spectrum and share it among themselves. The main objective of the dissertation is to investigate throughput limits and efficient methods to exploit multiuser diversity in cognitive radios. In a cognitive (secondary) network which is subject to interference power constraints imposed by a primary system, it is desirable to mitigate the interference on the primary *and* to harvest multiuser diversity gains in the secondary.

This problem of cognitive radio is often formulated as maximizing the secondary rate subject to interference constraints on the primary, or as the dual problem of minimizing the interference on the primary subject to a fixed rate for the secondary. Thus, reducing the interference footprint of the secondary is of paramount interest in spectrum sharing. Multihop relaying and cooperative communication is known to significantly mitigate interference and increase the sum-throughput in many multi-user scenarios [2], among others in broadcast channels [3], multiple access channels [4] and interference channels [5]. This has motivated the use of relays in spectrum sharing networks [6–12]. Another objective of the dissertation is to investigate efficient secondary relaying methods as well as the fundamental secondary throughput limits.

1.3 Contributions and Outline

Chapter 2 proposes a multiplicative superposition, and shows that it attains the optimal degrees of freedom under a wide set of antenna configurations and coherence lengths. Two signaling schemes are constructed based on the multiplicative superposition. In the first method, the messages of the two receivers are carried in the row and column spaces of a matrix, respectively. This method works better than orthogonal transmission while reception at each receiver is still interference-free. The second method uses coherent signaling for the receiver with CSIR, and Grassmannian signaling for the receiver without CSIR. This second method requires interference cancellation at the receiver with CSIR, but achieves higher DoF than the first method.

Chapter 3 extends product superposition to the domain of coherent signaling with pilots, demonstrates the advantages of product superposition in low-SNR as well as high-SNR, and established DoF optimality in a wider set of receiver antenna configurations. Two classes of decoders, with and without interference cancellation, are studied, and the effect of power allocation and partial CSI at the base station are also investigated.

Chapter 4 analyzes the sum throughput of a multiuser cognitive radio system with multi-antenna base stations, either in the uplink or downlink mode. The primary and secondary have N_p and n users, respectively, and their base stations have M_p and m antennas, respectively. We show that an *uplink secondary* throughput grows with $\frac{m}{N_p+1}\log n$ if the primary is a downlink system, and grows with $\frac{m}{M_p+1}\log n$ if the primary is an uplink system. These growth rates are shown to be optimal and can be obtained with a simple threshold-based user selection rule. In addition, we show that the secondary throughput can grow proportional to $\log n$ while simultaneously the interference on the primary is forced down to zero, asymptotically. For a *downlink secondary* it is shown that the throughput grows with $m \log \log n$ in the presence of either an uplink or downlink primary system. In addition, the interference on the primary can be made to go to zero asymptotically while the secondary throughput increases proportionally to $\log \log n$. The effect of unequal path loss and shadowing is also studied. It is shown that under a broad class of path loss and shadowing models, the secondary throughput growth rates remain unaffected.

Chapter 5 proposes a two-step (hybrid) scheduling method to harvest both interference diversity and secondary multiuser diversity. The method pre-selects a set of secondary users based on their interference on the primary, and from among them selects the user(s) that yield the highest secondary throughput. The optimal number of active secondary transmitters is characterized as a function of the primary interference constraint, the secondary transmit power, and the number of secondary transmitters n. The secondary sum-rate (throughput) of the proposed algorithm grows optimally (proportional to $\log n$). We investigate the tradeoff between scaling the secondary throughput and reducing interference on the primary, and characterize the optimum tradeoff in the regime of large n. Finally, we study user scheduling under fairness constraints, which is necessary when the channel statistics of secondary nodes are not identical. A modified hybrid scheduling rule is proposed to ensure user fairness, while still achieving the optimal growth rate for the secondary throughput.

Chapter 6 considers a spectrum-sharing network where n secondary relays are used to increase secondary rate and also mitigate interference on the primary by reducing the required overall secondary emitted power. We propose a distributed relay selection and clustering framework, obtain closed-form expressions for the secondary rate, and show that secondary rate increases proportionally to $\log n$. Remarkably, this is on the same order as the growth rate obtained in the *absence* of a primary system and its imposed constraints. To address the rate loss due to half-duplex relays, we propose an enhanced cognitive relaying protocol. Our results show that to maximize rate, the secondary relays must transmit with power proportional to n^{-1} (thus the sum of relay powers is bounded) and also that the secondary source may not operate at its maximum allowable power. Our results also characterize the tradeoff between the secondary rate and the interference on the primary, showing that the primary interference can be reduced asymptotically to zero as n increases, while still maintaining secondary rate that grows proportionally to $\log n$.

CHAPTER 2
GRASSMANNIAN PRODUCT SUPERPOSITION FOR MIMO BROADCAST CHANNELS

2.1 Introduction

In the MIMO broadcast channel, when channel state information is available at the receiver (CSIR) but not at the transmitter (CSIT), orthogonal transmission (e.g., TDMA) achieves optimal degrees of freedom (DoF) [13, 14]. With neither CSIT nor CSIR, again orthogonal transmission achieves the best possible DoF [15]. This chapter studies the broadcast channel where one receiver has full CSIR and another has no CSIR. In this case, new DoF gains are discovered that can be unlocked with novel signaling strategies.

The study of broadcast channels with unequal CSIR is motivated by downlink scenarios where users have different mobilities. Low-mobility users have the opportunity to reliably estimate their channels, while the high-mobility users may not have the same opportunity. For example, two mobiles operating at 2.1 GHz and moving at the speed of 5 km/h and 60 km/h have coherence times of approximately 50 ms and 4 ms, respectively [16]. If the transmitter broadcasts pilots every 50 ms, the low-mobility user is able to maintain accurate CSIR, while the high-mobility user cannot. This is partially due to the fact that effective estimation of the channel coefficients requires a training update every 4 ms, which induces significant overhead.

The main result of this chapter is that when one receiver has full CSIR and the other has none, the achieved DoF is strictly better than that obtained by orthogonal transmission. For the unequal CSIR scenario, we propose a *product superposition*, where the signals of the two receivers are multiplied to produce the broadcast signal.

In the following the receiver with full CSIR is referred to as the *static receiver* and the receiver with no CSIR as the *dynamic receiver.* Two classes of product superposition signaling are proposed:

- In the first method, information for both receivers is conveyed by the row and column spaces of a transmit signal matrix, respectively. The signal matrix is constructed from a product of two signals that lie on different Grassmannians. The two receivers do *not* interfere with each other even though there is no CSIT, a main point of departure from traditional superposition broadcasting [13, 17].
- In the second method, information for the static receiver is carried by the signal matrix values (coherent signaling), while information for the dynamic receiver is transported on the Grassmannian. The static receiver is required to decode and cancel interference, therefore this method is slightly more involved, but it achieves higher DoF compared with the first method.

Using the proposed methods, the exact DoF region is found when $N_d \leq N_s \leq M$, $T \geq 2N_d$, where N_d, N_s and M are the number of antennas at the dynamic receiver, static receiver and transmitter, respectively, and T is the channel coherence time of the dynamic receiver. For $N_s < N_d \leq M$, $T \geq 2N_d$, we partially characterize the DoF region when either the channel is the more capable type [18], or when the message set is degraded [19].

We use the following notation throughout the chapter: for a matrix $\mathbf{A}$, the transpose is denoted with $\mathbf{A}^t$, the conjugate transpose with $\mathbf{A}^H$, and the element in row i and column j with $[\mathbf{A}]_{i,j}$. The $k \times k$ identity matrix is denoted with $\mathbf{I}_k$. The set of $n \times m$ complex matrices is denoted with $\mathcal{C}^{n\times m}$.

The organization of this chapter is as follows. In Section 2.2 we introduce the system model and preliminary results. Two signaling methods are proposed and studied in Section 2.3 and Section 2.4, respectively.

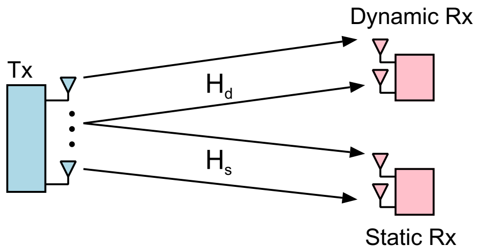

Figure 2.1. Channel model of Chapter 2.

2.2 System Model and Preliminaries

We consider a broadcast channel with an M-antenna transmitter and two receivers. One receiver has access to channel state information (CSI), and is referred to as the *static receiver.* The other receiver has no CSI, e.g. due to mobility, and is referred to as the *dynamic receiver.* The dynamic receiver has N_d antennas and the static receiver has N_s antennas. Denote the channel coefficient matrices from the transmitter to the dynamic and static receivers by $\mathbf{H}_d \in \mathcal{C}^{N_d \times M}$ and $\mathbf{H}_s \in \mathcal{C}^{N_s \times M}$, respectively. We assume that $\mathbf{H}_d$ is constant for T symbols (block-fading) and is unknown to both receivers, while $\mathbf{H}_s$ is known by the static receiver but not known by the dynamic receiver.[1] Neither $\mathbf{H}_d$ nor $\mathbf{H}_s$ is known by the transmitter (no CSIT).

Over T time-slots (symbols) the transmitter sends $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_M]^t$ across M antennas, where $\mathbf{x}_i \in \mathcal{C}^{T \times 1}$ is the signal vector sent by the antenna i. The normalized signal at the dynamic and static receivers is respectively

$$\begin{aligned} \mathbf{Y}_d &= \mathbf{H}_d \mathbf{X} + \frac{1}{\sqrt{\rho}} \mathbf{W}_d, \\ \mathbf{Y}_s &= \mathbf{H}_s \mathbf{X} + \frac{1}{\sqrt{\rho}} \mathbf{W}_s, \end{aligned} \tag{2.1}$$

[1] In practice $\mathbf{H}_s$ for a static receiver may vary across intervals of length much greater than T. However, for the purposes of this chapter, once $\mathbf{H}_s$ is assumed to be known to the static receiver, its time variation (or lack thereof) does not play any role in the subsequent mathematical developments. Therefore in the interest of elegance and for a minimal description of the requirements for the results, we only state that $\mathbf{H}_s$ is known.

where $\mathbf{W}_d \in \mathcal{C}^{N_d \times T}$ and $\mathbf{W}_s \in \mathcal{C}^{N_s \times T}$ are additive noise with i.i.d. entries $\mathcal{CN}(0,1)$. Each row of $\mathbf{Y}_d \in \mathcal{C}^{N_d \times T}$ (or $\mathbf{Y}_s \in \mathcal{C}^{N_s \times T}$) corresponds to the received signal at an antenna of the dynamic receiver (or the static receiver) over T time-slots. The transmitter is assumed to have an average power constraint ρ, and therefore, in the normalized channel model given by (2.1), the average power constraint is:

$$\mathbb{E}\big[\sum_{i=1}^{M} \mathrm{tr}(\mathbf{x}_i \mathbf{x}_i^H)\big] = T. \tag{2.2}$$

The channel $\mathbf{H}_d$ has i.i.d. entries with zero mean and unit variance, but we do *not* assign any specific distribution for $\mathbf{H}_d$. This general model includes Rayleigh fading as a special case where the entries of $\mathbf{H}_d$ are i.i.d. $\mathcal{CN}(0,1)$. The channel $\mathbf{H}_s$ is assumed to have full rank; this assumption, e.g., holds with probability 1 if the entries of $\mathbf{H}_s$ are drawn independently according to a continuous distribution. We focus on the case of $M = \max(N_d, N_s)$ and $T \geq 2N_d$, which is motivated by the fact that having more transmit antennas does not increase the multiplexing gain for either receiver, and the fact that if $T < 2N_d$, some of the antennas of the dynamic receiver can be deactivated without any loss in the degrees of freedom (DoF) [20].

The degrees of freedom at the dynamic and static receivers are defined as:

$$d_d = \lim_{\rho \to \infty} \frac{R_d(\rho)}{\log \rho}, \quad d_s = \lim_{\rho \to \infty} \frac{R_s(\rho)}{\log \rho},$$

where $R_d(\rho)$ and $R_s(\rho)$ are the rate of the dynamic receiver and the static receiver, respectively.

2.2.1 Definitions

Definition 2.2.1 (Isotropically Distributed Matrix [21]) *A random matrix $\mathbf{X} \in \mathcal{C}^{k \times n}$, where $n \geq k$, is called isotropically distributed (i.d.) if its distribution is invariant under unitary transformations, i.e., for any deterministic $n \times n$ unitary matrix $\mathbf{\Phi}$,*

$$p(\mathbf{X}) = p(\mathbf{X}\mathbf{\Phi}). \tag{2.3}$$

An example of i.d. matrices is $\mathbf{X}$ with i.i.d. $\mathcal{CN}(0,1)$ entries.

Remark 2.2.1 *An interesting property of i.d. matrices is that if* $\mathbf{X}$ *is i.d. and* $\mathbf{\Phi}$ *is a random unitary matrix that is independent of* $\mathbf{X}$*, then* $\mathbf{X\Phi}$ *is independent of* $\mathbf{\Phi}$ *[20, Lemma 4]. That is, any rotation to an i.d. matrix is essentially "invisible."*

Definition 2.2.2 (Stiefel manifold [22]) *The Stiefel manifold* $\mathbb{F}(n,k)$*, where* $n > k$*, is the set of all* $k \times n$ *unitary matrices, i.e.,*

$$\mathbb{F}(n,k) = \left\{\mathbf{Q} \in \mathcal{C}^{k\times n} : \mathbf{Q}\mathbf{Q}^H = \mathbf{I}_k\right\}.$$

For $k = 1$, the manifold $\mathbb{F}(n,1)$ is the collection of all n-dimensional vectors with unit norm, i.e., the surface of a unit ball.

Definition 2.2.3 (Grassmann manifold [22]) *The Grassmann manifold* $\mathbb{G}(n,k)$*, where* $n > k$*, is the set of all* k*-dimensional subspaces of* $\mathcal{C}^n$*.*

Remark 2.2.2 *The (complex) dimension of* $\mathbb{G}(n,k)$ *is*

$$\dim\big(\mathbb{G}(n,k)\big) = k(n-k), \tag{2.4}$$

i.e., each point in $\mathbb{G}(n,k)$ *has a neighborhood that is equivalent (homeomorphic) to a ball in the Euclidean space of complex dimension* $k(n-k)$*. The dimensionality of Grassmannian can also be viewed as follows. For any matrix* $\mathbf{Q}$*, there exists a* $k \times k$ *full rank matrix* $\mathbf{U}$ *so that*

$$\mathbf{Q}^* = \mathbf{U}\mathbf{Q} = \begin{bmatrix} 1 & \cdots & 0 & x_{1,k+1} & \cdots & x_{1n} \\ 0 & \cdots & 0 & & & \\ \vdots & & \vdots & & & \vdots \\ 0 & \cdots & 1 & x_{k,k+1} & \cdots & x_{kn} \end{bmatrix}, \tag{2.5}$$

where $\mathbf{Q}$ *and* $\mathbf{Q}^*$ *span the same row space. Therefore, each point in* $\mathbb{G}(n,k)$ *is determined by* $k(n-k)$ *complex parameters* x_{ji}*, for* $1 \leq j \leq k$ *and* $k+1 \leq i \leq n$*. In other words, a* k*-dimension subspace in* $\mathcal{C}^n$ *is uniquely decided by* $k(n-k)$ *complex variables.*

2.2.2 Non-coherent Point-to-point Channels

The analysis in this chapter uses insights and results from non-coherent communication in point-to-point MIMO channels, which are briefly outlined below.

Intuition

Consider a point-to-point $M \times N$ MIMO channel where the receiver does not know the channel $\mathbf{H}$, namely a non-coherent channel.

At high SNR the additive noise is negligible, so the received signal $\mathbf{Y} \approx \mathbf{HX}$, where $\mathbf{X}$ is the transmitted signal. Because $\mathbf{X}$ is multiplied by a random and unknown $\mathbf{H}$, the receiver cannot decode $\mathbf{X}$. However, communication is still possible because, for any non-singular $\mathbf{H}$, the received signal $\mathbf{Y}$ spans the same row space as $\mathbf{X}$. Therefore, the row space of $\mathbf{X}$ can be used to carry information without the need to know $\mathbf{H}$, i.e., the codebook consists of matrices with different row spaces.

Conveying information via subspaces can be viewed as communication on the Grassmann manifold where each distinct point in the manifold represents a different subspace [20]. In this case, the codewords (information) are represented by subspaces, which differs from the coherent communication that maps each codeword into one point in a Euclidean space [23]. Intuitively, the information of a Grassmannian codeword is carried by $k(n-k)$ variables, as seen in (2.5).

Optimal Signaling

The design of an optimal signaling can be viewed as sphere packing over Grassmannians [20]. At high SNR, the optimal signals are isotropically distributed unitary matrices [20, 21]. In addition, the optimal number of transmit antennas depends on the channel coherence time. For a short coherence interval, using fewer antennas may lead to a higher capacity. The optimal number of transmit antennas is

$$K = \min(M, N, \lfloor T/2 \rfloor), \tag{2.6}$$

where T is the channel coherence time, i.e., the number of symbols that the channel remains constant. Therefore, the optimal signals are $K \times T$ unitary matrices. In other words, K antennas ($K \leq M$) are in use and they transmit equal-energy and mutually orthogonal vectors. These unitary matrices reside in $\mathbb{G}(T, K)$ and each is interpreted as a representation of the subspace it spans. This method achieves the *maximum* DoF $K(T - K)$ over T time-slots. Note that the DoF coincides with the dimensionality of the Grassmannian $\mathbb{G}(T, K)$.

Subspace Decoding

Unlike coherent communication, in non-coherent signaling the information is embedded in the subspaces instead of the signal values. As long as two matrices span the same subspace, they correspond to the same message. Maximum-likelihood decoding chooses the codeword whose corresponding subspace is the closest one to the subspace spanned by the received signal. For example in [24], the received signals are projected on the subspaces spanned by different codewords, and then the one is chosen with the maximum projection energy. More precisely, for the transmitted signals $\mathbf{X}_i \in \mathcal{C}^{K\times T}$ from a unitary codebook $\mathcal{X}$, and the received signals $\mathbf{Y} \in \mathcal{C}^{K\times T}$, the ML detector is

$$\hat{\mathbf{X}}_{ML} = \arg\max_{\mathbf{X}_i \in \mathcal{X}} tr\{\mathbf{Y}\mathbf{X}_i^H \mathbf{X}_i \mathbf{Y}^H\}. \tag{2.7}$$

2.2.3 A Baseline Scheme: Orthogonal Transmission

For the purposes of establishing a baseline for comparison, we begin by considering a time-sharing (orthogonal transmission) that acquires CSIR via training in each interval and uses Gaussian signaling. This baseline method has been chosen to highlight the differences of the heterogeneous MIMO broadcast channel of this chapter with two other known scenarios: It is known that for a broadcast channel with no CSIT and perfect CSIR, orthogonal transmission ahieves the optimal DoF region [14]. Also, a training-based method with Gaussian signaling is sufficient to achieve DoF optimality [20] for the point-to-point noncoherent MIMO channel[2].

In orthogonal transmission, the transmitter communicates with the two receivers in a time-sharing manner. When transmitting to the dynamic receiver, it is optimal if the transmitter activates only K out of M antennas: it sends pilots from the K antennas sequentially over the first K time-slots; the dynamic receiver estimates the channel by using, e.g., minimum-mean-square-error (MMSE) estimation. Then, the transmitter sends data during the remaining $(T-K)$ time-slots, and the dynamic receiver decodes the data by using the estimated channel coefficients [20, 25]. Using this strategy, the maximum rate achieved by the dynamic receiver is:

$$K(1-\frac{K}{T})\log\rho+O(1). \tag{2.8}$$

The operating point in the achievable DoF region where the transmitter communicates exclusively with the dynamic receiver is denoted with $\mathcal{D}_1$.

$$\mathcal{D}_1=\big(K(1-\frac{K}{T}),\,0\big). \tag{2.9}$$

For the static receiver the channel is assumed to be known at the receiver, therefore data is transmitted to it coherently. The maximum rate achieved by the static receiver is [26]

$$\min(M,N_s)\log\rho+O(1). \tag{2.10}$$

The operating point in the DoF region where the transmitter communicates only with the static receiver is denoted with $\mathcal{D}_2$.

$$\mathcal{D}_2=\big(0,\,\min(M,N_s)\big). \tag{2.11}$$

Time-sharing between the two points of $\mathcal{D}_1$ and $\mathcal{D}_2$ yields the achievable DoF region

$$\bigg(tK(1-\frac{K}{T}),\,(1-t)\min(M,N_s)\bigg), \tag{2.12}$$

where t is a time-sharing variable.

[2] Grassmannian signaling is superior, but the same slope of the rate vs. SNR curve is obtained with training and Gaussian signaling in the point-to-point MIMO channel.

2.3 Grassmannian Superposition for Broadcast Channel

In this section, we propose a signaling method that attains DoF region superior to orthogonal transmission, and allows each receiver to decode its message while being oblivious of the other receiver's message.

2.3.1 A Toy Example

Consider $M = N_s = 2$, $N_d = 1$ and $T = 2$. From Section 2.2.3, orthogonal transmission attains 1/2 DoF per time-slot for the dynamic receiver and 2 DoF per time-slot for the static receiver. By time-sharing between the two receivers, the following DoF region is achieved

$$(\frac{t}{2},\ 2-2t), \tag{2.13}$$

where $t \in [0,1]$ is a time-sharing parameter.

We now consider the transmitter sends a product of signal vectors over 2 time-slots

$$\mathbf{X} = \mathbf{x}_s\mathbf{x}_d^t \in \mathcal{C}^{2\times 2}, \tag{2.14}$$

where $\mathbf{x}_d = [x_1^{(1)}\ x_2^{(1)}]^t$ and $\mathbf{x}_s = [x_1^{(2)}\ x_2^{(2)}]^t$ are the signals for the dynamic receiver and the static receiver, respectively. The vectors $\mathbf{x}_d$ and $\mathbf{x}_s$ have unit-norm and from codebooks that lie on $\mathbb{G}(2,1)$.

The signal at the dynamic receiver is

$$\begin{aligned}\mathbf{y}_d &= [h_1^{(1)}\ h_2^{(1)}]\begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \end{bmatrix}[x_1^{(1)}\ x_2^{(1)}] + \frac{1}{\sqrt{\rho}}[w_1^{(1)}\ w_2^{(1)}] \\ &= \tilde{h}^{(1)}\,[x_1^{(1)}\ x_2^{(1)}] + \frac{1}{\sqrt{\rho}}[w_1^{(1)}\ w_2^{(1)}],\end{aligned} \tag{2.15}$$

where $[h_1^{(1)}, h_2^{(1)}]$ is the isotropically distributed channel vector, and $\tilde{h}^{(1)}$ is the equivalent channel coefficient seen by the dynamic receiver.

The subspace spanned by $\mathbf{x}_d^t$ is the same as $\tilde{h}^{(1)}\mathbf{x}_d^t$, so at high SNR the dynamic receiver is able to determine the direction specified by $\mathbf{x}_d^t$. From Section 2.2.2, the dynamic receiver attains 1/2 DoF per time-slot, which is optimal even in the absence of the static receiver.

Consider the signal of the static receiver at time-slot 1:

$$\mathbf{y}_s = \mathbf{H}_s \begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \end{bmatrix} x_1^{(1)} + \frac{1}{\sqrt{\rho}} \begin{bmatrix} w_1^{(2)} \\ w_2^{(2)} \end{bmatrix}. \tag{2.16}$$

Because the static receiver knows $\mathbf{H}_s$, it can invert the channel as long as $\mathbf{H}_s$ is non-singular:

$$\left(\mathbf{H}_s^{-1}\mathbf{y}_s\right)^t = x_1^{(1)}[x_1^{(2)}\ x_2^{(2)}] + [w_1^{(2)}\ w_2^{(2)}]\mathbf{H}_s^{-t}. \tag{2.17}$$

The equivalent (unknown) channel seen by the static receiver is $x_1^{(1)}$, i.e., part of the dynamic receiver's signal. Using Grassmannian signaling via the subspace of $\mathbf{x}_s$, the DoF achieved is again 1/2 per time-slot.

Time-sharing between the proposed scheme and $\mathcal{D}_2$ (transmitting only to the static receiver) yields the achievable DoF region

$$\left(\frac{1}{2}t,\ 2 - \frac{3}{2}t\right). \tag{2.18}$$

The above region is strictly larger than that of orthogonal transmission, as shown in Figure 2.2. The static receiver achieves 1/2 DoF "for free" in the sense that this DoF was extracted for the static receiver without reducing the dynamic receiver's DoF.

2.3.2 Grassmannian Superposition Signaling

Based on the previous example, we design a general signaling method (the Grassmannian superposition) with two properties: (1) information is carried by subspaces and (2) two signal matrices are superimposed multiplicatively so that their row (or column) space is unaffected by multiplying the other receiver's signal matrix. Two separate cases are considered based on whether the number of static receiver antennas is larger than the number of dynamic receiver antennas.

Signaling In The Case $N_d < N_s$

The transmitter sends $\mathbf{X} \in \mathcal{C}^{N_s \times T}$ across $M = N_s$ antennas over an interval of length T:

$$\mathbf{X} = \sqrt{\frac{T}{N_d}}\mathbf{X}_s\mathbf{X}_d, \tag{2.19}$$

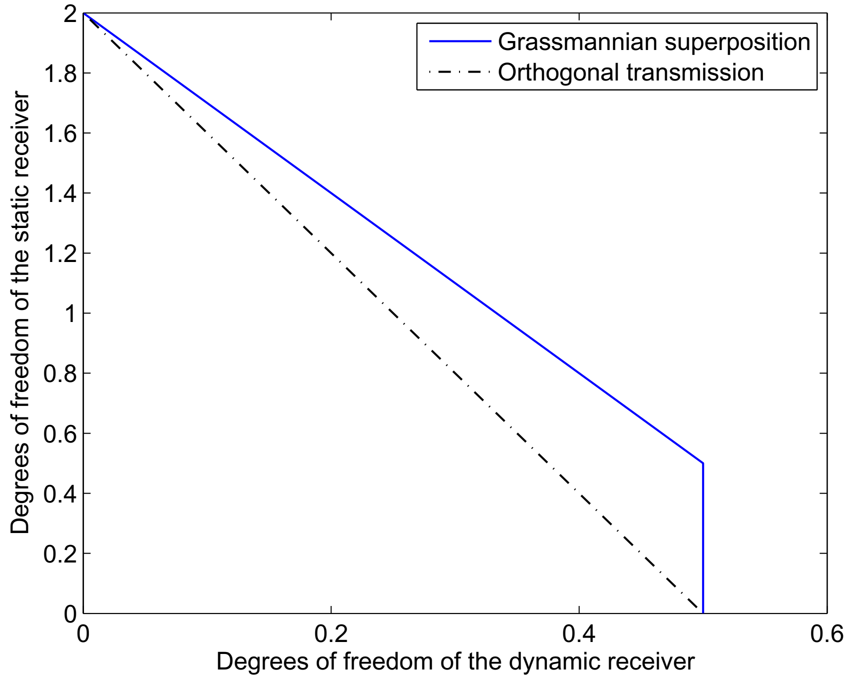

Figure 2.2. DoF region of the toy example 1

where $\mathbf{X}_d \in \mathcal{C}^{N_d \times T}$ and $\mathbf{X}_s \in \mathcal{C}^{N_s \times N_d}$ are the signals for the dynamic receiver and the static receiver, respectively. Here, $\sqrt{T/N_d}$ is a normalizing factor to satisfy the power constraint (2.2). Information for both receivers are sent over the Grassmannian, namely $\mathbf{X}_d$ is from a codebook $\mathcal{X}_d \subset \mathbb{G}(T, N_d)$ and $\mathbf{X}_s$ is from a codebook $\mathcal{X}_s \subset \mathbb{G}(N_s, N_d)$. The codebook $\mathcal{X}_d$ and $\mathcal{X}_s$ are chosen to be isotropically distributed unitary matrices (see Section 2.3.3 for more details).

A sketch of the argument for the DoF achieved by the Grassmannian superposition is as follows. The noise is negligible at high SNR, so the signal at the dynamic receiver is approximately

$$\mathbf{Y}_d \approx \sqrt{\frac{T}{N_d}}\mathbf{H}_d\mathbf{X}_s\mathbf{X}_d \in \mathcal{C}^{N_d \times T}. \tag{2.20}$$

The row space of $\mathbf{X}_d$ can be determined based on $\mathbf{Y}_d$, and then $(T - N_d)N_d$ independent variables (DoF) that specify the row space are recovered, i.e., the transmitted point $\mathbf{X}_d$ in $\mathcal{X}_d \in \mathbb{G}(T, N_d)$ is found.

For the static receiver, since $\mathbf{H}_s$ is known by the receiver, it inverts the channel (given that $\mathbf{H}_s$ is non-singular)

$$\mathbf{H}_s^{-1}\mathbf{Y}_s \approx \sqrt{\frac{T}{N_d}}\mathbf{X}_s\mathbf{X}_d \in \mathcal{C}^{N_s \times T}, \tag{2.21}$$

which has approximately the same column space as $\mathbf{X}_s$. The transmitted point $\mathbf{X}_s$ in $\mathcal{X}_s \in \mathbb{G}(N_s, N_d)$ will be recovered from the column space of $\mathbf{H}_s^{-1}\mathbf{Y}_s$, producing $(N_s - N_d)N_d$ DoF.

Therefore, the proposed scheme attains the DoF pair

$$\mathcal{D}_3 = \bigg(N_d(1 - \frac{N_d}{T}), \, \frac{N_d}{T}(N_s - N_d)\bigg). \tag{2.22}$$

The result is more formally stated as follows:

Theorem 2.3.1 **($N_d < N_s$)** *Consider a broadcast channel with an M-antenna transmitter, a dynamic receiver and a static receiver with N_d and N_s antennas, respectively, with coherence time T for the dynamic channel. The Grassmannian superposition achieves the rate pair*

$$\begin{cases} R_d = N_d\big(1 - \frac{N_d}{T}\big) \log \rho + O(1) \\ R_s = \frac{N_d}{T}(N_s - N_d) \log \rho + O(1) \end{cases}.$$

The corresponding DoF pair is denoted $\mathcal{D}_3 = \bigg(N_d(1 - \frac{N_d}{T}), \, \frac{N_d}{T}(N_s - N_d)\bigg)$. If we denote the DoF for the single-user operating points for the dynamic and static user with $\mathcal{D}_1$, $\mathcal{D}_2$ respectively, the achievable DoF region consists of the convex hull of $\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{D}_3$.

Proof See Section 2.5.1.

From Theorem 2.3.1 the static receiver attains a “free” rate of

$$\Delta R_1 = \frac{N_d}{T}(N_s - N_d) \log \rho + O(1). \tag{2.23}$$

We plot the achievable DoF region of Theorem 2.3.1 in Figure 2.3. For small T, the DoF gain achieved by the proposed method is significant, while as T increases, both methods approach the coherent upper bound [14] where both of the receivers have CSIR. For $T \to \infty$,

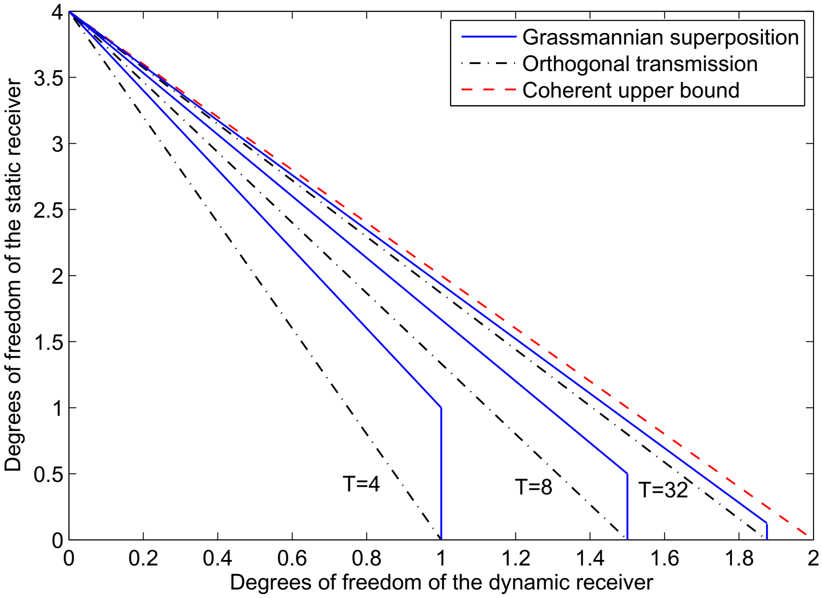

Figure 2.3. DoF region (Theorem 2.3.1): $N_d = 2$, $N_s = 4$.

the rate gain $\Delta R_1 = O(1)$, and no DoF gain is obtained. In this case, the achievable DoF region in Theorem 2.3.1 coincides with that attained by orthogonal transmission as well as the coherent outer bound [14]. This is not surprising, since if the channel remains constant ($T \to \infty$), the resource used for obtaining CSIR is negligible. Finally, the rate gain ΔR_1 is an increasing function of $(N_s - N_d)$, i.e., the extra antennas available for the static receiver.

Now, we design the dimension of $\mathbf{X}_d$ and $\mathbf{X}_s$ in (2.19) to maximize the achievable DoF region. To find the optimal dimensions, we allow the signaling to use a flexible number of antennas and time slots, up to the maximum available. Let $\mathbf{X}_d \in \mathcal{C}^{\hat{N}_d \times \hat{T}}$ and $\mathbf{X}_s \in \mathcal{C}^{\hat{N}_s \times \hat{N}_d}$, where $\hat{T} \leq T$, $\hat{N}_d \leq N_d$ and $\hat{N}_s \leq N_s$. Theorem 2.3.1 does not immediately reveal the optimal values of $\hat{N}_d$, $\hat{N}_s$, and $\hat{T}$, because the rates are not monotonic in the mentioned parameters. The following corollary presents the optimal value of $\hat{N}_d$, $\hat{N}_s$ and $\hat{T}$.

Corollary 2.3.2 *For the Grassmannian superposition under $N_d < N_s$, the signal dimension $\hat{T} = T$, $\hat{N}_d = N_d$ and $\hat{N}_s = N_s$ optimizes the achievable DoF region.*

Proof See Section 2.5.2.

Thus, in the special case of $N_d < N_s$, it is optimal to use all time slots and all antennas.

Signaling In The Case $N_d \geq N_s$

In this case, we shall see that sometimes the Grassmanian superposition may still outperform orthogonal transmission, but also under certain conditions (e.g. very large T or $N_d \gg N_s$) the Grassmannian superposition as described in this section may not improve the DoF compared with orthogonal transmission.

When $N_d \geq N_s$, if the Grassmannian signaling to the dynamic receiver uses all the N_d dimensions, there will remain no room for communication with the static receiver. To allow the static user to also use the channel, the dynamic user must "back off" from using all the rate available to it, in other words, the dimensionality of the signaling for the dynamic receiver must be reduced. The largest value of $\hat{N}_d$ that makes $\hat{N}_d < N_s$ and thus allows nontrivial Grassmannian superposition is $\hat{N}_d = N_s - 1$. Once we are in this regime, the results of the subsection 2.3.2 can be used. Specifically, Corollary 2.3.2 indicates that deactivating any further dynamic user antennas will not improve the DoF region. Thus, given N_s, and assuming we wish to have a non-trivial Grassmannian signaling for both users, using $\hat{N}_d = N_s - 1$ dimensions for signaling to the dynamic receiver maximizes the DoF region. The transmit signal is then

$$\mathbf{X} = \sqrt{\frac{T}{N_d}}\mathbf{X}_s\mathbf{X}_d, \tag{2.24}$$

where $\mathbf{X}_d \in \mathcal{C}^{(N_s-1)\times T}$ and $\mathbf{X}_s \in \mathcal{C}^{N_s\times(N_s-1)}$. The corresponding achievable DoF pair is

$$\mathcal{D}_4 = \Big((N_s-1)(1-\frac{N_s-1}{T}),\, (N_s-1)/T\Big), \tag{2.25}$$

which leads to the following result.

Corollary 2.3.3 ($N_d \geq N_s$) *Consider an M-antenna transmitter broadcasting to a dynamic receiver and a static receiver with N_d and N_s antennas, respectively, with coherence time T*

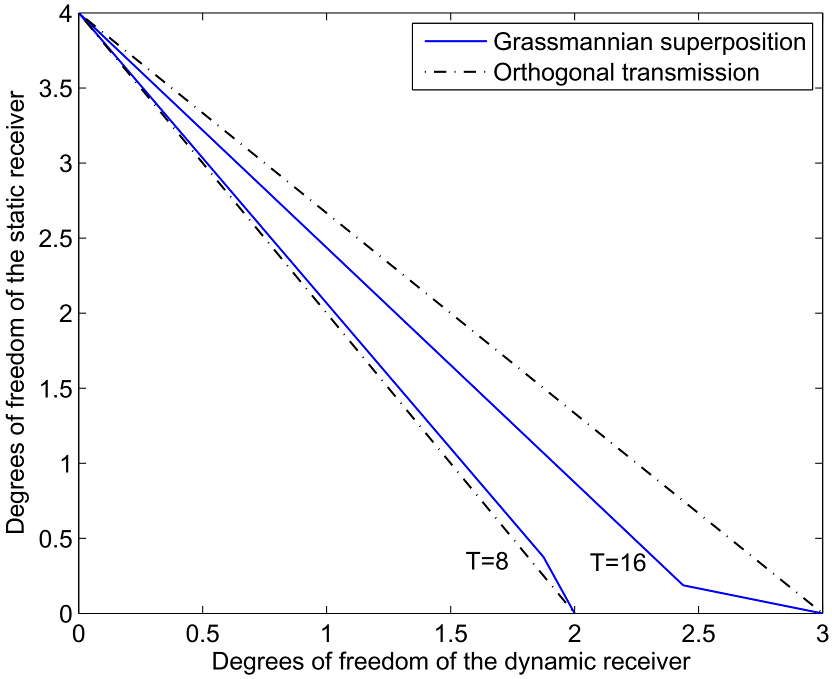

Figure 2.4. DoF region (Corollary 2.3.3): $N_d = N_s = 4$.

for the dynamic channel. Then the Grassmannian superposition achieves the rate pair

$$\begin{cases} R_d = (N_s - 1)\big(1 - \frac{N_s-1}{T}\big)\log\rho + O(1) \\ R_s = \frac{N_s-1}{T}\log\rho + O(1) \end{cases}.$$

Denote the corresponding DoF pair with $\mathcal{D}_4$. Together with the two single-user operating points $\mathcal{D}_1$ and $\mathcal{D}_2$ obtained earlier, the achievable DoF region consists of the convex hull of $\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{D}_4$.

Proof The proof follows directly by replacing N_d with $(N_s - 1)$ in Theorem 2.3.1.

In Corollary 2.3.3, the DoF for the static receiver has not been achieved for "free" but at the expense of reducing the DoF for the dynamic receiver. The transmitter uses only $N_s - 1$ dimensions for the dynamic receiver, which allows an extra DoF $(N_s - 1)/T$ to be attained at the static receiver. If $N_d - N_s$ and T are small, then the DoF gain of the static receiver outweighs the DoF loss for the dynamic, so that the overall achievable DoF region will be superior to that of orthogonal transmission. In contrast, if $N_d \gg N_s$ or T is large, the DoF loss from the dynamic receiver may not be compensated by the DoF gain from the static receiver, as illustrated by Figure 2.4. Therefore in the latter case orthogonal transmission may do better. The following corollary specifies the condition under which Grassmannian superposition improves DoF region compared with orthogonal transmission.

Corollary 2.3.4 *For $N_d \geq N_s$, the Grassmannian superposition improves DoF region with respect to orthogonal transmission if and only if*

$$\frac{N_s - (N_s - 1)/T}{\big(N_s - 1)(1 - (N_s - 1)/T\big)} < \frac{N_s}{N_d(1 - N_d/T)}. \tag{2.26}$$

Proof The necessary and sufficient condition for ensuring the improvement of the achievable DoF region is that the slope between $\mathcal{D}_2$ and $\mathcal{D}_4$ is larger than the slope between $\mathcal{D}_1$ and $\mathcal{D}_2$, which is equivalent to the inequality in the corollary.

2.3.3 Design of $\mathcal{X}_d$ and $\mathcal{X}_s$

The representation of a point in the Grassmannian is not unique [27] (also see Remark 2.2.2), and therefore the codebooks $\mathcal{X}_d \subset \mathbb{G}(T, N_d)$ and $\mathcal{X}_s \subset \mathbb{G}(N_s, N_d)$ are not unique.

First, $\mathcal{X}_s$ is chosen to be a unitary codebook. When $\mathbf{X}_s$ is unitary, for i.i.d. Rayleigh fading $\mathbf{H}_d$, the equivalent dynamic channel $\widetilde{\mathbf{H}}_d = \mathbf{H}_d\mathbf{X}_s$ still has i.i.d. Rayleigh fading coefficients [21]. Therefore, the static receiver is *transparent* to the dynamic receiver, which allows us to decouple and simplify the design of the two codebooks and their decoders.

Once $\mathcal{X}_s$ is chosen to be a set of unitary matrices, communication between dynamic receiver and the transmitter is equivalent to a non-coherent point-to-point MIMO channel. Hence, to maximize the rate of the dynamic receiver at high SNR, $\mathcal{X}_d$ must also be a collection of isotropically distributed unitary matrices (see Section 2.2).

Remark 2.3.1 *With unitary codebooks $\mathcal{X}_d$ and $\mathcal{X}_s$, information for both receivers is conveyed purely by the the subspace to which the codeword belongs. Consider $\mathbf{X} \in \mathcal{C}^{k\times n}$, $n \geq k$, which* *is uniquely represented by Ω (the row space of $\mathbf{X}$) and a $k \times k$ coefficient matrix $\mathbf{C}$ according to a certain basis of Ω. The codewords $\mathbf{X}_d, \mathbf{X}_s$ can be represented as*

$$\begin{aligned} \mathbf{X}_d &\rightarrow (\Omega_d, \mathbf{C}_d), \\ \mathbf{X}_s &\rightarrow (\Omega_s, \mathbf{C}_s). \end{aligned} \tag{2.27}$$

In a manner similar to [20], one can verify

$$I(\mathbf{X}_d; \mathbf{Y}_d) = I(\Omega_d; \mathbf{Y}_d) + \underbrace{I(\mathbf{C}_d; \mathbf{Y}_d | \Omega_d)}_{=0}, \tag{2.28}$$

and

$$I(\mathbf{X}_s; \mathbf{Y}_s | \mathbf{H}_s) = I(\Omega_s; \mathbf{Y}_s | \mathbf{H}_s) + \underbrace{I(\mathbf{C}_s; \mathbf{Y}_s | \Omega_s, \mathbf{H}_s)}_{=0}. \tag{2.29}$$

2.3.4 Multiplicative vs. Additive Superposition

In this section, we compare product superposition with additive superposition. Under additive superposition, the transmit signal has a general expression

$$\mathbf{X} = \sqrt{c_1 \rho}\, \mathbf{V}_1\, \mathbf{X}_d + \sqrt{c_2 \rho}\, \mathbf{V}_2\, \mathbf{X}_s, \tag{2.30}$$

where $\mathbf{V}_1$ and $\mathbf{V}_2$ are the precoding matrices, and c_1 and c_2 represent the power allocation. In this case, the signal at the dynamic receiver is

$$\mathbf{Y}_d = \sqrt{c_1 \rho}\, \mathbf{H}_d \mathbf{V}_1 \mathbf{X}_d + \sqrt{c_2 \rho}\, \mathbf{H}_d \mathbf{V}_2 \mathbf{X}_s + \mathbf{W}_d. \tag{2.31}$$

Since $\mathbf{H}_d$ is unknown, the second interference term cannot be completely eliminated in general, which leads to a bounded signal-to-interference-plus-noise ratio (SINR), resulting in zero DoF for the dynamic receiver.

For the multiplicative superposition, the signal at the dynamic receiver is

$$\begin{aligned} \mathbf{Y}_d &= \sqrt{c\rho}\, \mathbf{H}_d \mathbf{X}_s \mathbf{X}_d + \mathbf{W}_d \\ &= \sqrt{c\rho}\, \widetilde{\mathbf{H}}_d \mathbf{X}_d + \mathbf{W}_d, \end{aligned} \tag{2.32}$$

where c is a power normalizing constant. For any unitary $\mathbf{X}_s$, $\mathbf{X}_s\mathbf{X}_d$ and $\mathbf{X}_d$ span the same row space. This invariant property of Grassmannian enables us to convey information to the static receiver via $\mathbf{X}_s$ without reducing the degrees of freedom of the dynamic receiver. Intuitively, the dynamic receiver does not have CSIR and is "insensitive" to rotation, i.e., the distribution of $\mathbf{Y}_d$ does not depend on $\mathbf{X}_s$.

For the static receiver, the received signal is

$$\mathbf{Y}_s = \sqrt{c\rho}\,\mathbf{H}_s\mathbf{X}_s\mathbf{X}_d + \mathbf{W}_s. \tag{2.33}$$

Because $\mathbf{H}_s$ is known, the channel rotation $\mathbf{X}_s$ is detectable, i.e., the distribution of $\mathbf{Y}_s$ depends on $\mathbf{X}_s$. Therefore $\mathbf{X}_s$ can be used to convey information for the static receiver.

2.4 Grassmannian-Euclidean Superposition for the Broadcast Channel

We now propose a new transmission scheme based on successive interference cancellation, where the static receiver decodes and removes the signal for the dynamic receiver before decoding its own signal. This scheme improves the DoF region compared to the non-interfering Grassmannian signaling of the previous section.

2.4.1 A Toy Example

Consider $M = N_d = N_s = 1$ and $T = 2$. Our approach is that over 2 time-slots, the transmitter sends

$$\mathbf{X} = x_s\,\mathbf{x}_d^t \in \mathcal{C}^{1\times 2}, \tag{2.34}$$

where $\mathbf{x}_d = [x_1^{(1)}\ x_2^{(1)}]^t$ is the signal for the dynamic receiver and x_s is the signal for the static receiver. Here, $\mathbf{x}_d$ has unit-norm and is from a codebook $\mathcal{X}_d$ that is a subset of $\mathbb{G}(2,1)$, and x_s can obey any distribution that satisfies the average power constraint.

The signal at the dynamic receiver is

$$\mathbf{y}_d = h_d x_s [x_1^{(1)} \ x_2^{(1)}] + \frac{1}{\sqrt{\rho}} [w_1^{(1)} \ w_2^{(1)}] \tag{2.35}$$

$$= \tilde{h}_d \, [x_1^{(1)} \ x_2^{(1)}] + \frac{1}{\sqrt{\rho}} [w_1^{(1)} \ w_2^{(1)}], \tag{2.36}$$

where h_d is the channel coefficient of the dynamic receiver, and $\tilde{h}_d \triangleq h_d x_s$ is the equivalent channel coefficient. The dynamic receiver can determine the row space spanned by $\mathbf{x}_d$ even though $\tilde{h}_d$ is unknown, in a manner similar to Section 2.3.1. The total DoF conveyed by $\mathbf{x}_d$ is 1 (thus $\frac{1}{2}$ per time-slot); this is the optimal DoF under the same number of antennas and coherence time.

For the static receiver, the received signal is:

$$\mathbf{y}_s = h_s x_s [x_1^{(1)} \ x_2^{(1)}] + \frac{1}{\sqrt{\rho}} [w_1^{(2)} \ w_2^{(2)}] \tag{2.37}$$

$$= \tilde{h}_s \, [x_1^{(1)} \ x_2^{(1)}] + \frac{1}{\sqrt{\rho}} [w_1^{(2)} \ w_2^{(2)}], \tag{2.38}$$

where h_s is the channel coefficient of the static receiver, and $\tilde{h}_s \triangleq h_s x_s$. Intuitively, since (2.36) and (2.38) are equivalent, if the dynamic receiver decodes the subspace of $\mathbf{x}_d$, so does the static receiver. Then, the exact signal vector $\mathbf{x}_d$ is known to the static receiver (recall that each subspace is uniquely represented by a signal matrix). The static receiver removes the interference signal $\mathbf{x}_d$

$$\mathbf{y}_s \mathbf{x}_d^H = h_s x_s + \frac{1}{\sqrt{\rho}} \tilde{w}_s, \tag{2.39}$$

where $\tilde{w}_s$ is the equivalent noise. Finally, the static receiver knows h_s, so it decodes x_s and attains 1/2 DoF per time-slot.

Therefore, the proposed scheme attains the maximum DoF for the dynamic receiver, meanwhile achieving 1/2 DoF for the static receiver. With time sharing between this scheme and $\mathcal{D}_2$, the achievable DoF pair is

$$(d_d, d_s) = \left(\frac{t}{2}, 1 - \frac{t}{2}\right). \tag{2.40}$$

Figure 2.5 shows that this region is uniformly larger than that of orthogonal transmission.

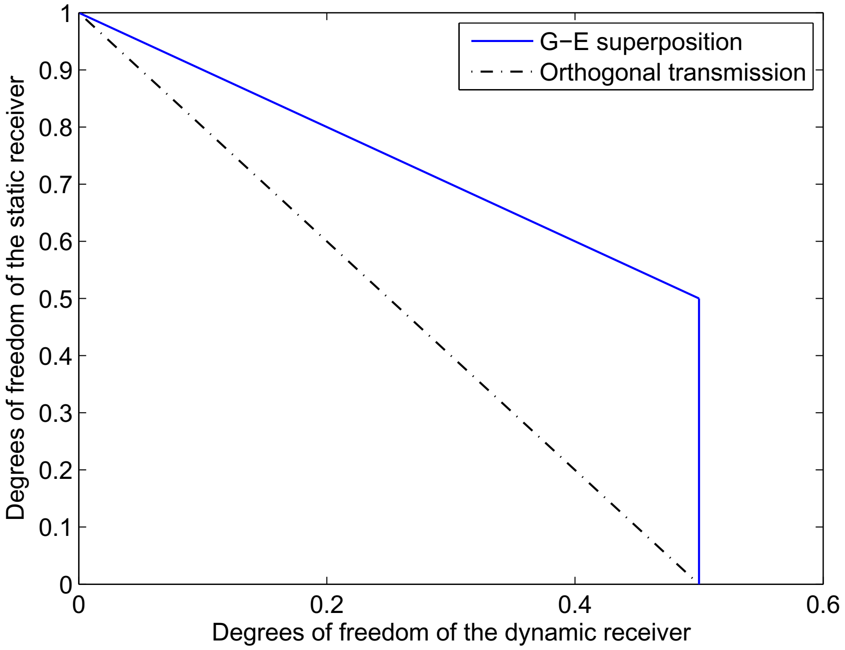

Figure 2.5. DoF region of the toy example 2.

Remark 2.4.1 *There are two key differences between the method proposed here and the Grassmannian superposition proposed in Section 2.3. First, the information for the static receiver is carried by the* value *of x_s instead of its direction (subspace), i.e., the signal for the static receiver is carried in the Euclidean space. Second, the static receiver must decode and remove the interference signal for the dynamic receiver before decoding its own signal, which is unlike the non-interfering method of the previous section.*

2.4.2 Grassmannian-Euclidean Superposition Signaling

We denote the aforementioned method as *Grassmannian-Euclidean superposition*, whose generalization is the subject of this subsection. Two separate cases are considered based on whether the number of static receiver antennas is less than, or no less than, the number of dynamic receiver antennas.

Signaling In The Case $N_d \leq N_s$

The transmitter sends $\mathbf{X} \in \mathcal{C}^{N_s \times T}$

$$\mathbf{X} = \sqrt{\frac{T}{N_d N_s}} \mathbf{X}_s \mathbf{X}_d, \tag{2.41}$$

where $\mathbf{X}_d \in \mathcal{C}^{N_d \times T}$ and $\mathbf{X}_s \in \mathcal{C}^{N_s \times N_d}$ are signals for the dynamic receiver and the static receiver, respectively. The signal $\mathbf{X}_d$ is from a Grassmannian codebook $\mathcal{X}_d \subset \mathbb{G}(T, N_d)$, while $\mathbf{X}_s$ is from a conventional Gaussian codebook $\mathcal{X}_s$. The constant $\sqrt{T/N_d N_s}$ is a power normalizing factor.

We now give a sketch of the argument of the DoF attained by the superposition signaling (2.41). For the dynamic receiver, $\mathbf{Y}_d \approx \mathbf{H}_d \mathbf{X}_s \mathbf{X}_d$ at high SNR. When $N_d \leq N_s$, the equivalent channel $\mathbf{H}_d \mathbf{X}_s \in \mathcal{C}^{N_d \times N_d}$ has full rank and does not change the row space of $\mathbf{X}_d$. Recovering the row space of $\mathbf{X}_d$ produces $(T - N_d)N_d$ DoF, which is similar to Section 2.3.

For the static receiver, the signal at high SNR is

$$\mathbf{Y}_s \approx \sqrt{\frac{T}{N_d N_s}} \mathbf{H}_s \, \mathbf{X}_s \, \mathbf{X}_d = \sqrt{\frac{T}{N_d N_s}} \widetilde{\mathbf{H}}_s \, \mathbf{X}_d. \tag{2.42}$$

For $N_d \leq N_s$, $\widetilde{\mathbf{H}}_s = \mathbf{H}_s \mathbf{X}_s \in \mathcal{C}^{N_s \times N_d}$ has full column rank and does not change the the row space of $\mathbf{X}_d$, and therefore, the signal intended for the dynamic receiver can be decoded by the static receiver. From the subspace spanned by $\mathbf{X}_d$, the codeword $\mathbf{X}_d \in \mathcal{X}_d$ is identified. Then, $\mathbf{X}_d$ is peeled off from the static signal:

$$\mathbf{Y}_s \mathbf{X}_d^H \approx \sqrt{\frac{T}{N_d N_s}} \mathbf{H}_s \mathbf{X}_s \in \mathcal{C}^{N_s \times N_d}. \tag{2.43}$$

Because $\mathbf{H}_s$ is known by the static receiver, Eq. (2.43) is a point-to-point MIMO channel. Therefore, $N_s N_d$ DoF can be communicated via $\mathbf{X}_s$ to the static receiver (over T time-slots) [26].

Altogether, the Grassmannian-Euclidean superposition attains the DoF pair $\mathcal{D}_5$

$$\mathcal{D}_5 = \Big(N_d(1 - N_d/T), \, N_s N_d / T \Big). \tag{2.44}$$

More precisely, we have the following theorem.

Theorem 2.4.1 **($N_d \leq N_s$)** *Consider a broadcast channel with an M-antenna transmitter, a dynamic receiver and a static receiver with N_d and N_s antennas, respectively, with coherence time T for the dynamic channel. The Grassmannian-Euclidean superposition achieves the rate pair*

$$\begin{cases} R_d = N_d\big(1 - \frac{N_d}{T}\big)\log\rho + O(1) \\ R_s = \frac{N_d N_s}{T}\log\rho + O(1) \end{cases}.$$

Denote the corresponding DoF pair by $\mathcal{D}_5$. Together with the two single-user operating points $\mathcal{D}_1$, $\mathcal{D}_2$ obtained earlier, the achievable DoF region consists of the convex hull of $\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{D}_5$.

Proof See Section 2.5.3.

With the Grassmannian-Euclidean superposition, the static receiver attains the following gain compared with orthogonal transmission:

$$\Delta R_2 = \frac{N_d N_s}{T}\log\rho + O(1). \tag{2.45}$$

From Figure 2.6, for relatively small T or large N_s, the DoF gain is significant. For example, at $T = 2N_d$, the minimum coherence interval considered in this chapter, the proposed method achieves a DoF $N_s/2$ for the static receiver while attaining the maximum DoF $N_d/2$ for the dynamic receiver. As T increases the gain over orthogonal transmission decreases. In the limit $T \to \infty$, we have $\Delta R_2 = O(1)$, and the DoF gain of Grassmannian-Euclidean superposition goes away. The Grassmannian-Euclidean superposition also provides DoF gain over the non-interfering Grassmannian superposition[3]

$$\Delta R = \frac{N_d^2}{T}\log\rho + O(1). \tag{2.46}$$

The optimal design of the dimensions of $\mathbf{X}_d$ and $\mathbf{X}_s$ is trivial, because the DoF region in Theorem 2.4.1 is indeed optimal (see Section 2.4.4).

[3] Although Grassmannian-Euclidean superposition achieves larger DoF than the non-interfering Grassmannian superposition, it may not achieve larger rate at low or moderate SNR due to the decodable restriction on the rate (interference).

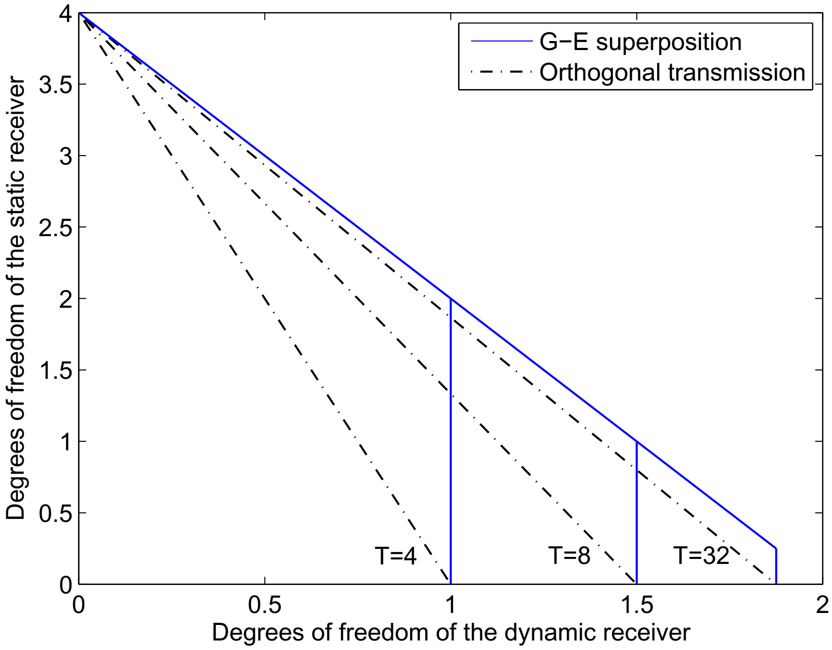

Figure 2.6. DoF region (Theorem 2.4.1): $N_d = 2$, $N_s = 4$.

Signaling In The Case $N_d > N_s$

When the static receiver has fewer antennas than the dynamic receiver, it may not be able to decode the dynamic signal. Here, we cannot directly apply the signaling structure given by (2.41). A straightforward way is to activate only N_s antennas at the transmitter and use only N_s dimensions for the dynamic receiver, that is

$$\mathbf{X} = \sqrt{\frac{T}{N_s^2}}\mathbf{X}_s\mathbf{X}_d \in \mathcal{C}^{N_s \times T}, \tag{2.47}$$

where $\mathbf{X}_d \in \mathcal{C}^{N_s \times T}$ and $\mathbf{X}_s \in \mathcal{C}^{N_s \times N_s}$, and $\sqrt{T/N_s^2}$ is a power normalizing factor.

Following the same argument as the case of $N_d \leq N_s$, the Grassmannian-Euclidean superposition achieves the DoF pair

$$\mathcal{D}_6 = \left(N_s(1 - \frac{N_s}{T}),\, \frac{N_s^2}{T}\right). \tag{2.48}$$

Corollary 2.4.2 ($N_d > N_s$) *Consider a broadcast channel with an M-antenna transmitter, a dynamic receiver and a static receiver with N_d and N_s antennas, respectively, with coherence time T for the dynamic channel. The Grassmannian-Euclidean superposition achieves the rate pair*

$$\begin{cases} R_d = N_s\big(1 - \frac{N_s}{T}\big)\log\rho + O(1) \\ R_s = \frac{N_s^2}{T}\log\rho + O(1) \end{cases}$$

Denote the corresponding DoF pair with $\mathcal{D}_6$. Together with the two single-user operating points $\mathcal{D}_1$ and $\mathcal{D}_2$ obtained earlier, the achievable DoF region consists of the convex hull of $\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{D}_6$.

Proof The proof directly follows from Theorem 2.4.1.

In Corollary 2.4.2, the static rate receiver is obtained at the expense of a reduction in the dynamic rate. The transmitter uses only N_s out of N_d dimensions available for the dynamic receiver, which allows extra DoF N_s^2/T for the static receiver. A necessary and sufficient condition for Grassmannian-Euclidean superposition to improve the DoF region is as follows.

Corollary 2.4.3 *For the Grassmannian-Euclidean superposition, the signal dimension $\hat{T} = T$, $\hat{N}_d = N_d$ and $\hat{N}_s = N_s$ optimizes the rate region at high SNR. Moreover, it achieves superior DoF region compared with orthogonal transmission if and only if*

$$N_s > (1 - \frac{N_d}{T})N_d \tag{2.49}$$

Proof First, using the maximum number of static antennas ($\hat{N}_s = N_s$) is optimal, because both R_d and R_s in Corollary 2.4.2 are increasing functions of N_s (note that $N_s \leq T/2$).

Second, we find the optimal $\hat{T}$. Maximizing the achievable DoF region is equivalent to maximizing the slope of the line between $\mathcal{D}_2$ and $\mathcal{D}_4$, i.e.,

$$(0,\, N_s) \quad \text{and} \quad \big(N_s(1 - \frac{N_s}{\hat{T}}),\, \frac{N_s^2}{\hat{T}}\big), \tag{2.50}$$

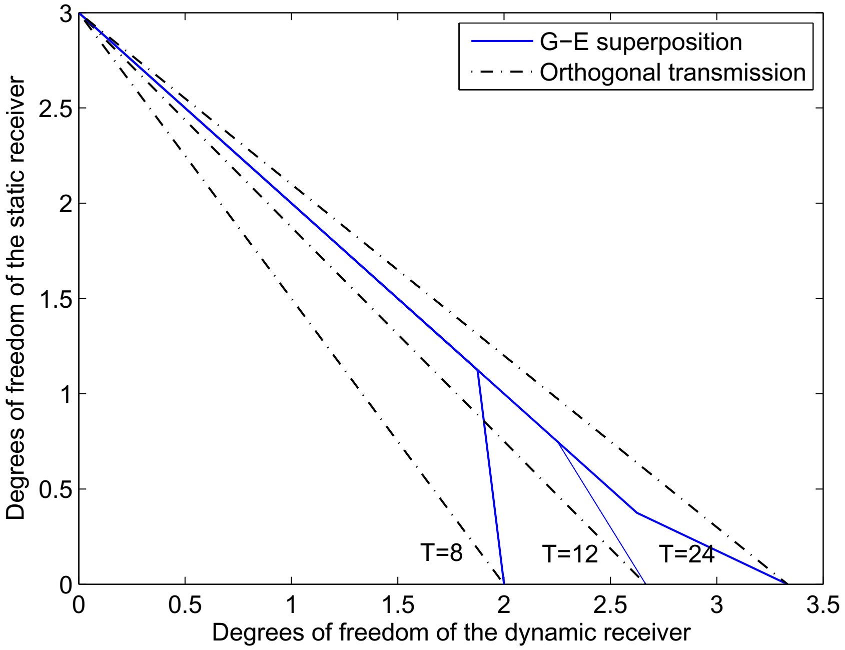

Figure 2.7. DoF region (Corollary 2.4.2): $N_d = 4$, $N_s = 3$.

which has a constant slope -1 and is independent of $\hat{T}$. Therefore, any choice of $\hat{T}$, as long as $\hat{T} \geq 2N_s$, achieves a boundary point of the DoF region of the Grassmannian-Euclidean superposition.

Finally, for the Grassmannian-Euclidean superposition to be superior to orthogonal transmission in term of DoF, the slope of the line between $\mathcal{D}_2$ and $\mathcal{D}_6$ must be larger than the slope between $\mathcal{D}_1$ and $\mathcal{D}_2$, namely

$$\frac{N_s}{(1 - N_d/T)N_d} > 1. \tag{2.51}$$

This completes the proof.

Corollary 2.4.3 can be interpreted as follows: the Grassmannian-Euclidean superposition achieves superior DoF if and only if the maximum DoF of the static receiver is larger than that of the dynamic receiver.

2.4.3 Design of $\mathcal{X}_d$ and $\mathcal{X}_s$

We heuristically argue that it is reasonable to choose $\mathcal{X}_d$ to be isotropically distributed unitary matrices and $\mathcal{X}_s$ to be i.i.d. complex Gaussian codebook. .

Recall that the Grassmannian-Euclidean superposition is to allow the static receiver to decode the signal for the dynamic receiver and then remove this interference. After interference cancellation, the static receiver has an equivalent point-to-point MIMO channel with perfect CSIR, in which case Gaussian signal achieves capacity.

Assuming $\mathbf{X}_s \in \mathcal{X}_s$ has i.i.d. $\mathcal{CN}(0,1)$ entries, the equivalent channel for the dynamic receiver $\mathbf{H}_d\mathbf{X}_s$ is isotropically distributed (see Definition 2.2.1), which leads to two properties. First, for any $T \times T$ unitary matrix $\mathbf{\Phi}$,

$$p(\mathbf{Y}_d\mathbf{\Phi} \,|\, \mathbf{X}_d\mathbf{\Phi}) = p(\mathbf{Y}_d \,|\, \mathbf{X}_d). \tag{2.52}$$

Second, for any $N_d \times N_d$ unitary matrix $\mathbf{\Psi}$

$$p(\mathbf{Y}_d \,|\, \mathbf{\Psi}\mathbf{X}_d) = p(\mathbf{Y}_d \,|\, \mathbf{X}_d). \tag{2.53}$$

Based on these properties, the optimal signaling structure for the channel of the dynamic receiver is a diagonal matrix[4] times a unitary matrix [20, 21]. Therefore, choosing $\mathcal{X}_d$ to be isotropically distributed unitary matrices is not far from optimal.

2.4.4 Degrees of Freedom Region

In this section, we show that the Grassmannian-Euclidean superposition achieves the optimal DoF region under certain channel conditions.

Degrees of Freedom In The Case $N_d \leq N_s$

In this case, the optimal DoF region is as follows.

[4] When the channel is i.i.d. Rayleigh fading this diagonal matrix should be identity at high SNR [20]. However, it remains unknown whether the optimal choice is an identity matrix at arbitrary SNR.

Corollary 2.4.4 **($N_d \leq N_s$)** *When an M-antenna transmitter transmits to a dynamic receiver and a static receiver with N_d and N_s antennas, respectively, with the dynamic channel coherence time T, the DoF region is:*

$$\begin{cases} d_d \leq N_d(1 - \frac{N_d}{T}) \\ \frac{d_d}{N_d} + \frac{d_s}{N_s} \leq 1 \end{cases} . \tag{2.54}$$

Proof An outer bound can be found when both receivers have CSIR. The DoF region of the coherent upper bound is [14]

$$\frac{d_d}{N_d} + \frac{d_s}{N_s} \leq 1. \tag{2.55}$$

An inner bound is attained by Grassmannian-Euclidean superposition, which reaches the boundary of (2.55) except for $d_d > N_d(1 - N_d/T)$. However, the DoF of the dynamic receiver can never exceed $N_d(1-N_d/T)$ (see Section 2.2). Therefore, Grassmannian-Euclidean superposition achieves the DoF region.

Degrees of Freedom In The Case $N_d > N_s$

In this case, the Grassmannian-Euclidean superposition does not match the coherent outer bound (2.55), however, we can partially characterize the DoF region for broadcasting with degraded message sets [19] and in the case of the more capable channel [18]. For both cases the capacity region is characterized by:

$$\begin{cases} R_d & \leq I(\mathbf{U}; \mathbf{Y}_d) \\ R_d + R_s & \leq I(\mathbf{X}_s; \mathbf{Y}_s|\mathbf{U}) + I(\mathbf{U}; \mathbf{Y}_d) \\ R_d + R_s & \leq I(\mathbf{X}_s; \mathbf{Y}_s) \end{cases} , \tag{2.56}$$

where $\mathbf{U}$ is an auxiliary random variable. From the last inequality we have

$$R_d + R_s \leq N_s \log \rho + O(1), \tag{2.57}$$

that is

$$d_d + d_s \leq N_s. \tag{2.58}$$

When $N_s \geq (1 - N_d/T)N_d$, the inner bound in Corollary 2.4.2 coincides with the outer bound (2.58) for $0 \leq d_d \leq N_s(1 - N_s/T)$, therefore, the DoF is established for this range. For $d_d > N_s(1 - N_s/T)$, the inner and outer bounds do not match, but the gap is small when N_s is close to N_d.

When $N_s < (1 - N_d/T)N_d$, the inner bound in Corollary 2.4.2 is inferior to orthogonal transmission and the problem remains open.

2.5 Proof of Theorem and Corollary

2.5.1 Proof of Theorem 2.3.1

Achievable Rate for the Dynamic Receiver

The normalized received signal $\mathbf{Y}_d \in \mathcal{C}^{N_d \times T}$ at the dynamic receiver is

$$\mathbf{Y}_d = \sqrt{\frac{T}{N_d}}\mathbf{H}_d\mathbf{X}_s\mathbf{X}_d + \frac{1}{\sqrt{\rho}}\mathbf{W}_d, \tag{2.59}$$

where $\mathbf{H}_d \in \mathcal{C}^{N_d \times N_s}$ is the dynamic channel, $\mathbf{X}_d \in \mathcal{C}^{N_d \times T}$ and $\mathbf{X}_s \in \mathcal{C}^{N_s \times N_d}$ are the isotropically distributed, unitary signals for the dynamic and static receivers, respectively, and $\mathbf{W}_d \in \mathcal{C}^{N_d \times T}$ is additive Gaussian noise.

Let $\tilde{\mathbf{H}}_d \triangleq \mathbf{H}_d\mathbf{X}_s$ be the $N_d \times N_d$ equivalent channel, and rewrite (2.59) as

$$\mathbf{Y}_d = \sqrt{\frac{T}{N_d}}\tilde{\mathbf{H}}_d\mathbf{X}_d + \frac{1}{\sqrt{\rho}}\mathbf{W}_d. \tag{2.60}$$

The elements in $\tilde{\mathbf{H}}_d$ are

$$\tilde{h}_{ij} = [\tilde{\mathbf{H}}_d]_{i,j} = \sum_{k=1}^{N_s} h_{ik}x_{kj}, \quad 1 \leq i, j \leq N_d, \tag{2.61}$$

where $h_{ik} = [\mathbf{H}_d]_{ik}$ and $x_{kj} = [\mathbf{X}_s]_{kj}$. Note that h_{ik} is i.i.d. random variable with zero mean and unit variance, therefore,

$$\mathbb{E}[\tilde{h}_{ij}^H\tilde{h}_{mn}] = 0, \qquad (i, j) \neq (m, n). \tag{2.62}$$

For $(i,j)=(m,n)$ we have

$$\mathbb{E}[|\tilde{h}_{ij}|^2] = \sum_{k=1}^{N_s} \mathbb{E}\big[|h_{ik}|^2 |x_{kj}|^2\big] \tag{2.63}$$

$$= \mathbb{E}\big[\sum_{k=1}^{N_s} |x_{kj}|^2\big] = 1, \tag{2.64}$$

where (2.64) holds because $\mathbb{E}[|h_{ik}|^2]=1$ and each column of $\mathbf{X}_s$ has unit norm. Therefore, the equivalent channel $\tilde{\mathbf{H}}_d$ has uncorrelated entries with zero mean and unit variance.

We now find a lower bound for the mutual information

$$I(\mathbf{X}_d;\mathbf{Y}_d) = h(\mathbf{Y}_d) - h(\mathbf{Y}_d|\mathbf{X}_d), \tag{2.65}$$

i.e., an achievable rate for the dynamic receiver. First, we find an upper bound for $h(\mathbf{Y}_d|\mathbf{X}_d)$. Let $\mathbf{y}_{di}$ be the row i of $\mathbf{Y}_d$. Using the independence bound on entropy:

$$h(\mathbf{Y}_d|\mathbf{X}_d) \le \sum_{i=1}^{N_d} h(\mathbf{y}_{di}|\mathbf{X}_d). \tag{2.66}$$

Let $\tilde{\mathbf{h}}_i$ be the row i of $\tilde{\mathbf{H}}_d$. Then, conditioned on $\mathbf{X}_d$ the covariance of $\mathbf{y}_{di}$ is

$$\mathbb{E}[\mathbf{y}_{di}^H\mathbf{y}_{di}|\mathbf{X}_d] = \frac{T}{N_d}\mathbf{X}_d^H\,\mathbb{E}\big[\tilde{\mathbf{h}}_i^H\tilde{\mathbf{h}}_i\big]\,\mathbf{X}_d + \frac{1}{\rho}\,\mathbf{I}_T \tag{2.67}$$

$$= \frac{T}{N_d}\mathbf{X}_d^H\mathbf{X}_d + \frac{1}{\rho}\,\mathbf{I}_T, \tag{2.68}$$

where the last equality holds since all the elements in $\tilde{\mathbf{H}}_d$ are uncorrelated with zero mean and unit variance. In addition, given $\mathbf{X}_d$, the vector $\mathbf{y}_{di}$ has zero mean, and therefore, $h(\mathbf{y}_{di}|\mathbf{X}_d)$ is upper bounded by the differential entropy of a multivariate normal random vector with the same covariance [17]:

$$h(\mathbf{y}_{di}|\mathbf{X}_d) \le \log\det\big(\frac{T}{N_d}\mathbf{X}_d^H\mathbf{X}_d + \frac{1}{\rho}\mathbf{I}\big) \tag{2.69}$$

$$\le N_d \log\big(\frac{T}{N_d} + \frac{1}{\rho}\big) - (T-N_d)\log\rho. \tag{2.70}$$

Combining (2.66) and (2.70), we obtain

$$h(\mathbf{Y}_d|\mathbf{X}_d) \le N_d^2 \log\big(\frac{T}{N_d} + \frac{1}{\rho}\big) - N_d(T-N_d)\log\rho. \tag{2.71}$$

After calculating the upper bound for $h(\mathbf{Y}_d|\mathbf{X}_d)$, we now find a lower bound for $h(\mathbf{Y}_d)$ as follows.

$$h(\mathbf{Y}_d) > h\big(\sqrt{\frac{T}{N_d}}\mathbf{H}_d\mathbf{X}_s\mathbf{X}_d\big) \tag{2.72}$$

$$\geq h\big(\sqrt{\frac{T}{N_d}}\mathbf{H}_d\mathbf{X}_s\mathbf{X}_d \,\big|\, \mathbf{H}_d, \mathbf{X}_s\big), \tag{2.73}$$

where (2.72) holds since we remove the noise, and (2.73) holds since conditioning does not increase differential entropy. The Jacobian from $\mathbf{X}_d$ to $\mathbf{H}_d\mathbf{X}_s\mathbf{X}_d$ is [28, Theorem. 2.1.5]:

$$J_{X_d} = \Big(\sqrt{\frac{T}{N_d N_s}}\det(\mathbf{H}_d\mathbf{X}_s)\Big)^{N_d}. \tag{2.74}$$

Therefore, from (2.73) we have

$$h(\mathbf{Y}_d) > h(\mathbf{X}_d) + \mathbb{E}[\log J_{X_d}], \tag{2.75}$$

where the expectation is with respect to $\mathbf{X}_s$ and $\mathbf{H}_d$. Because $\mathbf{X}_d$ is an isotropically distributed unitary matrix, i.e., uniformly distributed on the Stiefel manifold $\mathbb{F}(T, N_d)$, we have [20]

$$h(\mathbf{X}_d) = \log\big|\mathbb{F}(T, N_d)\big|, \tag{2.76}$$

where $\big|\mathbb{F}(T, N_d)\big|$ is the volume of $\mathbb{F}(T, N_d)$ based on the Haar measure induced by the Lebesgue measure restricted to the Stiefel manifold [28]:

$$\big|\mathbb{F}(T, N_d)\big| = \prod_{i=T-N_d+1}^{T} \frac{2\pi^i}{(i-1)!}. \tag{2.77}$$

Finally, combining (2.71) and (2.75), we obtain

$$I(\mathbf{X}_d; \mathbf{Y}_d) > N_d(T-N_d)\log\rho + \log|\mathbb{F}(T, N_d)| + \mathbb{E}[\log J_{X_d}] - N_d\sum_{i=1}^{N_d}\log\big(\frac{T}{N_d} + \frac{1}{\rho}\big) \tag{2.78}$$

$$= N_d(T-N_d)\log\rho + O(1). \tag{2.79}$$

Normalizing $I(\mathbf{X}_d; \mathbf{Y}_d)$ over T time-slots yields the achievable rate of the dynamic receiver.

Achievable Rate for the Static Receiver

The signal received at the static receiver is

$$\mathbf{Y}_s = \sqrt{\frac{T}{N_d}}\mathbf{H}_s\mathbf{X}_s\mathbf{X}_d + \frac{1}{\sqrt{\rho}}\mathbf{W}_s, \tag{2.80}$$

where $\mathbf{H}_s \in \mathcal{C}^{N_s \times N_s}$ is the static channel and $\mathbf{W}_s \in \mathcal{C}^{N_s \times T}$ is additive Gaussian noise. Denote the sub-matrix containing the first N_d columns of $\mathbf{Y}_s$ with $\mathbf{Y}'_s$.

$$\mathbf{Y}'_s = \sqrt{\frac{T}{N_d}}\mathbf{H}_s\mathbf{X}_s\mathbf{X}'_d + \frac{1}{\sqrt{\rho}}\mathbf{W}'_s, \tag{2.81}$$

where $\mathbf{X}'_d \in \mathcal{C}^{N_d \times N_d}$ is the corresponding sub-matrix of $\mathbf{X}_d$, and $\mathbf{W}'_s \in \mathcal{C}^{N_d \times N_d}$ is i.i.d. Gaussian noise. Given $\mathbf{H}_s$, the mutual information between $\mathbf{Y}_s$ and $\mathbf{X}_s$ is lower bounded by:

$$I(\mathbf{Y}_s; \mathbf{X}_s|\mathbf{H}_s) \geq I(\mathbf{Y}'_s; \mathbf{X}_s|\mathbf{H}_s). \tag{2.82}$$

We will focus on $I(\mathbf{Y}'_s; \mathbf{X}_s|\mathbf{H}_s)$ to derive a lower bound. Using the singular value decomposition (SVD):

$$\mathbf{H}_s = \mathbf{U}^H\mathbf{\Sigma}\mathbf{V}, \tag{2.83}$$

where $\mathbf{U}$, $\mathbf{V} \in \mathcal{C}^{N_s \times N_s}$ and

$$\mathbf{\Sigma} = \mathrm{diag}(\lambda_1, \cdots, \lambda_{N_s}) \tag{2.84}$$

with $|\lambda_1| \geq \cdots \geq |\lambda_{N_s}|$. Since $\mathbf{H}_s$ is known and non-singular, the dynamic receiver applies $\mathbf{H}_s^{-1}$ to remove it:

$$\mathbf{H}_s^{-1}\mathbf{Y}'_s = \sqrt{\frac{T}{N_d}}\mathbf{X}_s\mathbf{X}'_d + \frac{1}{\sqrt{\rho}}\mathbf{W}''_s. \tag{2.85}$$

The columns of $\mathbf{W}''_s$ are mutually independent, and each column has an autocorrelation:

$$\mathbf{R}_W = \mathbf{V}^H\,\mathbf{\Sigma}^{-2}\mathbf{V}. \tag{2.86}$$

Because mutual information is independent of the choice of coordinates, we have

$$I(\mathbf{Y}'_s; \mathbf{X}_s|\mathbf{H}_s) = I(\mathbf{H}_s^{-1}\mathbf{Y}'_s; \mathbf{X}_s|\mathbf{H}_s) \tag{2.87}$$

$$= h(\mathbf{H}_s^{-1}\mathbf{Y}'_s|\mathbf{H}_s) - h(\mathbf{H}_s^{-1}\mathbf{Y}'_s|\mathbf{X}_s, \mathbf{H}_s). \tag{2.88}$$

Let $\mathbf{y}_{si}$ be the column i of $\mathbf{H}_s^{-1}\mathbf{Y}'_s$, then via the independence bound on entropy:

$$h(\mathbf{H}_s^{-1}\mathbf{Y}'_s|\mathbf{X}_s, \mathbf{H}_s) \leq \sum_{i=1}^{N_d} h(\mathbf{y}_{si}|\mathbf{X}_s, \mathbf{H}_s). \tag{2.89}$$

From (2.85) and (2.86), the autocorrelation of $\mathbf{y}_{si}$ conditioned on $\mathbf{X}_s$ and $\mathbf{H}_s$ is

$$\begin{aligned} \mathbf{R}_{si} &= \frac{T}{N_d}\mathbf{X}_s\mathbb{E}[\mathbf{x}'_{di}\mathbf{x}'^H_{di}]\mathbf{X}_s^H + \frac{1}{\rho}\mathbf{R}_W & (2.90) \\ &= \frac{T}{N_d}\mathbf{X}_s\mathbf{R}_{di}\mathbf{X}_s^H + \frac{1}{\rho}\mathbf{R}_W & (2.91) \end{aligned}$$

where $\mathbf{x}'_{d,i} \in \mathcal{C}^{N_d\times 1}$ is the column i of $\mathbf{X}'_d$ and has autocorrelation $\mathbf{R}_{d,i}$. The expected value of $\mathbf{y}_{s,i}$ is zero and thus the differential entropy is maximized if $\mathbf{y}_{s,i}$ has multivariate normal distribution [17]:

$$\begin{aligned} h(\mathbf{y}_{si}|\mathbf{X}_s, \mathbf{H}_s) &\leq \log\det\big(\frac{T}{N_d}\mathbf{X}_s\mathbf{R}_{di}\mathbf{X}_s^H + \frac{1}{\rho}\mathbf{R}_W\big) \\ &= \log\det\big(\frac{T}{N_d}\mathbf{V}\mathbf{X}_s\mathbf{R}_{di}\mathbf{X}_s^H\mathbf{V}^H + \frac{1}{\rho}\mathbf{\Sigma}^{-2}\big). & (2.92) \end{aligned}$$

The following lemma calculates $\mathbf{R}_{di}$, the autocorrelation of a column of an i.d. matrix.

Lemma 2.5.1 *If $\mathbf{Q} \in \mathcal{C}^{T\times T}$ is isotropically distributed (i.d.) unitary matrix, then each row and column of $\mathbf{Q}$ is an i.d. unit vector with autocorrelation $\frac{1}{T}\mathbf{I}_T$.*

Proof From Definition 2.2.1, given $\mathbf{Q}$ is i.d., for any deterministic unitary matrix $\mathbf{\Phi} \in \mathcal{C}^{T\times T}$,

$$p(\mathbf{Q\Phi}) = p(\mathbf{Q}), \tag{2.93}$$

which implies that the marginal distribution of each row and column remains unchanged under any transform $\mathbf{\Phi}$. Therefore, each row and column is an i.d. unit vector. Without loss of generality, we consider the first row of $\mathbf{Q}$, denoted as $\mathbf{q}_1$. Let the autocorrelation of $\mathbf{q}_1$ be $\mathbf{R}_q$ and posses the eigenvalue decomposition $\mathbf{R}_q = \mathbf{P}^H\mathbf{\Sigma}_q\mathbf{P}$, where $\mathbf{P} \in \mathcal{C}^{n\times n}$ is unitary and $\mathbf{\Sigma}_q$ is diagonal. Since $\mathbf{q}_1\mathbf{P}^H$ has the same distribution as $\mathbf{q}_1$, therefore

$$\mathbf{R}_q = \mathbb{E}[\mathbf{q}_1^H\mathbf{q}_1] = \mathbf{P}\,\mathbb{E}[\mathbf{q}_1^H\mathbf{q}_1]\,\mathbf{P}^H = \mathbf{\Sigma}_q. \tag{2.94}$$

Thus $\mathbf{R}_q$ is a diagonal matrix. Furthermore, the diagonal elements of $\mathbf{\Sigma}_q$ have to be identical, i.e., $\mathbf{\Sigma}_q = a\mathbf{I}_T$, otherwise $\mathbf{R}_q$ would not be rotationally invariant which conflicts with an i.d. assumption. Finally, because $\mathrm{tr}(\mathbf{R}_q) = 1$, we have $\mathbf{\Sigma}_q = \mathbf{I}_T/T$. This completes the proof of Lemma 2.5.1.

Since $\mathbf{X}_d$ is an i.d. unitary matrix, based on Lemma 2.5.1, the autocorrelation of its sub-column is

$$\mathbf{R}_{di} = \mathbf{I}_{N_d}/T. \tag{2.95}$$

Therefore, the eigenvalues of $\mathbf{V}_2\mathbf{X}_s\mathbf{R}_{1i}\mathbf{X}_s^H\mathbf{V}_2^H$ are

$$\Bigg(\underbrace{\frac{1}{T}, \cdots, \frac{1}{T}}_{N_d}, \underbrace{0, \cdots, 0}_{N_s - N_d}\Bigg). \tag{2.96}$$

We now bound the eigenvalues of the sum of two matrices in (2.92), noting that λ_j^{-2} are in ascending order and using a theorem of Weyl [29, Theorem 4.3.1]:

$$h(\mathbf{y}_{si}|\mathbf{X}_s, \mathbf{H}_s) \leq N_d \log\big(\frac{1}{N_d} + \lambda_{N_s}^{-2}\big) + (N_s - N_d)\log\frac{1}{\rho}\lambda_{N_s}^{-2}. \tag{2.97}$$

From (2.89) and (2.97), we have:

$$h(\mathbf{H}_s^{-1}\mathbf{Y}_s'|\mathbf{X}_s, \mathbf{H}_s) \leq N_d^2 \log\big(\frac{1}{N_d} + \lambda_{N_s}^{-2}\big) + N_d(N_s - N_d)\log\lambda_{N_s}^{-2} - N_d(N_s - N_d)\log\rho. \tag{2.98}$$

We now calculate a lower bound for $h(\mathbf{H}_s^{-1}\mathbf{Y}_s'|\mathbf{H}_s)$:

$$h(\mathbf{H}_s^{-1}\mathbf{Y}_s'|\mathbf{H}_s) > h(\sqrt{\frac{T}{N_d}}\mathbf{X}_s\mathbf{X}_d'|\mathbf{H}_s) \tag{2.99}$$

$$> h(\sqrt{\frac{T}{N_d}}\mathbf{X}_s\mathbf{X}_d'|\mathbf{X}_d', \mathbf{H}_s). \tag{2.100}$$

From [28, Theorem. 2.1.5], given $\mathbf{X}_d'$ the Jacobian of the transformation from $\mathbf{X}_s$ to $\sqrt{\frac{T}{N_d}}\mathbf{X}_s\mathbf{X}_d'$ is:

$$J_{X_s} = \left(\sqrt{\frac{T}{N_d}}\right)^{N_s} \det(\mathbf{X}_d')^{N_d}. \tag{2.101}$$

Therefore, from the right hand side of (2.100) we have

$$h(\mathbf{H}_s^{-1}\mathbf{Y}_s'|\mathbf{H}_s) > h(\mathbf{X}_s) + \mathbb{E}[\log J_{X_s}], \tag{2.102}$$

where the expectation is with respect to $\mathbf{X}_d'$. Because $\mathbf{X}_s$ is uniformly distributed on the Stiefel manifold $\mathbb{F}(N_s, N_d)$, we have [20]

$$h(\mathbf{X}_d) = \log \big|\mathbb{F}(N_s, N_d)\big|, \tag{2.103}$$

where $\big|\mathbb{F}(N_s, N_d)\big|$ is the volume of $\mathbb{F}(N_s, N_d)$, which is given by [28]:

$$\big|\mathbb{F}(N_s, N_d)\big| = \prod_{i=N_s-N_d+1}^{N_s} \frac{2\pi^i}{(i-1)!}. \tag{2.104}$$

Finally, substituting (2.102) and (2.98) into (2.88), we have

$$I(\mathbf{Y}_s'; \mathbf{X}_s|\mathbf{H}_s) = N_d(N_s - N_d)\log\rho + O(1). \tag{2.105}$$

Hence, the rate achieved by the static receiver is

$$\frac{1}{T}\mathbb{E}[I(\mathbf{Y}_s'; \mathbf{X}_s|\mathbf{H}_s)] = \frac{N_d}{T}(N_s - N_d)\log\rho + O(1), \tag{2.106}$$

where the expectation is with respect to $\mathbf{H}_s$.

2.5.2 Proof of Corollary 2.3.2

The objective is to find the best dimensions for the transmit signals $\mathbf{X}_d \in \mathcal{C}^{\hat{N}_d \times \hat{T}}$ and $\mathbf{X}_s \in \mathcal{C}^{\hat{N}_s \times \hat{N}_d}$. From Theorem 2.3.1, it is easily determined that $\hat{N}_s = N_s$ is optimal, because the pre-log factor of R_2 increases with $\hat{N}_s$ and the pre-log factor of R_d is independent of $\hat{N}_s$ (given $\hat{N}_d \leq N_s$).

To find the optimal values of $\hat{N}_d, \hat{T}$, we start by relaxing the variables by allowing them to be continuous valued, i.e. $\hat{N}_d \to x$ and $\hat{T} \to y$, and then showing via the derivatives that the cost functions are monotonic, therefore optimal values reside at the boundaries, which are indeed integers.

Using the DoF expression from Theorem 2.3.1, the slope between two achievable points $\mathcal{D}_2$ and $\mathcal{D}_3$ is:

$$f(x,y) = \frac{x(N_s - x)/y - N_s}{x(1 - x/y)}. \tag{2.107}$$

Therefore, for all $0 < x \leq N_d$,

$$\frac{\partial f(x,y)}{\partial y} = \frac{x}{(y-x)^2} > 0. \tag{2.108}$$

We wish to maximize f with the constraint $y \leq T$, thus $y = T$ is optimal.

Substituting $y = T$ into $f(x,y)$, we have

$$\frac{\partial f(x,T)}{\partial x} = -\frac{(T - N_s)x^2 + TN_s x - T^2 N_s}{x^2(T-x)^2}. \tag{2.109}$$

If $T = N_s$, since $x \leq T/2$, then $\frac{\partial f}{\partial x} > 0$. In this case $x = N_d$ maximizes the DoF region.

If $T \neq N_s$, let $T = \alpha N_s$. When $0 < \alpha < \frac{3}{4}$, one can verify that $\frac{\partial f}{\partial x} > 0$ for all $x > 0$. Thus, $x = N_d$ is optimal. When $\alpha \geq \frac{3}{4}$, let $\frac{\partial f}{\partial x} = 0$, and we have the corresponding solutions:

$$x_{1,2} = \frac{-\alpha N_s \pm \alpha N_s \sqrt{1 + 4(\alpha - 1)}}{2(\alpha - 1)}. \tag{2.110}$$

When $\frac{3}{4} \leq \alpha < 1$, the above solutions are positive, where the smaller one is:

$$x_1 = \frac{\alpha N_s - \alpha N_s \sqrt{1 - 4(1-\alpha)}}{2(1-\alpha)} > N_d. \tag{2.111}$$

Since $\frac{\partial f}{\partial x} > 0$ at $x = 0$, we have $\frac{\partial f}{\partial x} > 0$ for $0 \leq x \leq N_d$. When $\alpha > 1$, the (only) positive solution of (2.110) is:

$$x_1 = \frac{\alpha N_s + \alpha N_s \sqrt{1 + 4(\alpha - 1)}}{2(\alpha - 1)} > N_d. \tag{2.112}$$

Once again, since $\frac{\partial f}{\partial x} > 0$ at $x = 0$, we have $\frac{\partial f}{\partial x} > 0$ for $0 \leq x \leq N_d$.

Therefore, for all cases, $x = N_d$ maximizes the DoF region.

2.5.3 Proof of Theorem 2.4.1

Achievable Rate for the Dynamic Receiver

The proof is similar to the proof for Theorem 2.3.1, so we only outline key steps. The received signal at the dynamic receiver is

$$\mathbf{Y}_d = \sqrt{\frac{T}{N_d N_s}}\mathbf{H}_d\mathbf{X}_s\mathbf{X}_d + \frac{1}{\sqrt{\rho}}\mathbf{W}_d, \tag{2.113}$$

where $\mathbf{Y}_d \in \mathcal{C}^{N_d\times T}$ and $\mathbf{H}_d \in \mathcal{C}^{N_d\times N_s}$ and $\mathbf{W}_d \in \mathcal{C}^{N_d\times T}$ is additive Gaussian noise. We establish a lower bound for the mutual information between $\mathbf{X}_d$ and $\mathbf{Y}_d$:

$$I(\mathbf{X}_d;\mathbf{Y}_d) = h(\mathbf{Y}_d) - h(\mathbf{Y}_d|\mathbf{X}_d). \tag{2.114}$$

In the above equation, we have

$$h(\mathbf{Y}_d|\mathbf{X}_d) \leq \sum_{i=1}^{N_d} h(\mathbf{y}_{di}|\mathbf{X}_d).$$

One can verify

$$h(\mathbf{y}_{di}|\mathbf{X}_d) \leq \log\det\big(\frac{T}{N_s}\mathbf{X}_d^H\mathbf{X}_d + \frac{1}{\rho}\mathbf{I}\big). \tag{2.115}$$

Finally, we obtain

$$h(\mathbf{Y}_d|\mathbf{X}_d) < N_d^2\log\big(\frac{T}{N_s} + \frac{1}{\rho}\big) - N_d(T-N_d)\log\rho. \tag{2.116}$$

The lower bound is given by:

$$h(\mathbf{Y}_d) > \log|\mathbb{F}(T,N_d)| + \mathbb{E}[\log J_{X_d}], \tag{2.117}$$

where the expectation is with respect to $\mathbf{H}_d$ and $\mathbf{X}_s$, and

$$J_{X_1} = \left(\sqrt{\frac{T}{N_d N_s}}\det(\mathbf{H}_d\mathbf{X}_s)\right)^{N_d}. \tag{2.118}$$

Combining (2.116) and (2.118), and normalizing over T time-slots leads to the achievable rate of the dynamic receiver.

Achievable Rate for the Static Receiver

The received signal at the static receiver is $\mathbf{Y}_s \in \mathcal{C}^{N_s \times T}$

$$\mathbf{Y}_s = \sqrt{\frac{T}{N_d N_s}} \mathbf{H}_s \mathbf{X}_s \mathbf{X}_d + \frac{1}{\sqrt{\rho}} \mathbf{W}_s,$$

where $\mathbf{H}_s \in \mathcal{C}^{N_s \times N_s}$ is the static channel, and $\mathbf{W}_s \in \mathcal{C}^{N_d \times T}$ is additive Gaussian noise.

We first calculate the *decodable* dynamic rate at the static receiver in the next lemma.

Lemma 2.5.2 *The static receiver is able to decode the dynamic rate R_d if*

$$R_d \leq N_d(1 - N_d/T) \log \rho + O(1). \tag{2.119}$$

Proof Use the SVD for $\mathbf{H}_s$ and re-write the signal at the static receiver as

$$\mathbf{Y}_s = \sqrt{\frac{T}{N_d N_s}} \mathbf{U}^H \mathbf{\Sigma} \mathbf{V} \mathbf{X}_s \mathbf{X}_d + \frac{1}{\sqrt{\rho}} \mathbf{W}_s, \tag{2.120}$$

Because $\mathbf{X}_s$ is an isotropically distributed unitary matrix, $\mathbf{X}'_s \triangleq \mathbf{V}\mathbf{X}_s$ has the same distribution as $\mathbf{X}_s$, i.e., a matrix of i.i.d. $\mathcal{CN}(0,1)$. Rotate $\mathbf{Y}_s$ with $\mathbf{U}$

$$\mathbf{Y}'_s \triangleq \mathbf{U}\mathbf{Y}_s = \sqrt{\frac{T}{N_d N_s}} \mathbf{\Sigma} \mathbf{X}'_s \mathbf{X}_d + \frac{1}{\sqrt{\rho}} \mathbf{W}'_s, \tag{2.121}$$

where $\mathbf{W}'_s$ is i.i.d. Gaussian noise. Let $\mathbf{Y}''_s \in \mathcal{C}^{N_d \times T}$ be the first N_d rows of $\mathbf{Y}'_s$, i.e., the rows corresponding to the largest N_d singular modes of $\mathbf{H}_s$, that is $|\lambda_1| \geq \cdots \geq |\lambda_{N_d}|$. We denote the corresponding $N_d \times N_d$ sub-matrix of $\mathbf{X}'_s$ by $\mathbf{X}''_s$. Then,

$$\mathbf{Y}''_s = \mathrm{diag}(\lambda_1, \cdots, \lambda_{N_d})\, \mathbf{X}''_s \mathbf{X}_d + \frac{1}{\sqrt{\rho}} \mathbf{W}''_s. \tag{2.122}$$

Conditioned on $\mathbf{H}_s$, the decodable dynamic rate at the static receiver is

$$I(\mathbf{X}_d; \mathbf{Y}_s | \mathbf{H}_s) = I(\mathbf{X}_d; \mathbf{Y}'_s | \mathbf{H}_s),$$

which is lower bounded by

$$I(\mathbf{X}_d; \mathbf{Y}''_s | \mathbf{H}_s) = h(\mathbf{Y}''_s | \mathbf{H}_s) - h(\mathbf{Y}''_s | \mathbf{X}_d, \mathbf{H}_s). \tag{2.123}$$

Using the independence bound for $h(\mathbf{Y}_s''|\mathbf{X}_d, \mathbf{H}_s)$ yields

$$h(\mathbf{Y}_s''|\mathbf{X}_d, \mathbf{H}_s) \leq \sum_{i=1}^{N_d} h(\mathbf{y}_{si}|\mathbf{X}_d, \mathbf{H}_s), \tag{2.124}$$

where $\mathbf{y}_{si}$ is the row i of $\mathbf{Y}_s''$. Let $\mathbf{x}_{si}$ be the row i of $\mathbf{X}_s''$, for $1 \leq i \leq N_d$. Since $\mathbf{X}_s'' \in \mathcal{C}^{N_d \times N_d}$ have i.i.d. $\mathcal{CN}(0,1)$ entries, all the row vectors $\mathbf{x}_{si}$ have the same autocorrelation I_{N_d}.

Conditioned on $\mathbf{X}_d$, the autocorrelation of $\mathbf{y}_{si} = \lambda_i \mathbf{x}_{si} \mathbf{X}_d$ is given by

$$\mathbb{E}[\mathbf{y}_{si}^H \mathbf{y}_{si}|\mathbf{X}_d, \mathbf{H}_s] = \lambda_i^2 \mathbf{X}_d^H \mathbf{X}_d + \frac{1}{\rho}\mathbf{I}_T. \tag{2.125}$$

Therefore,

$$h(\mathbf{y}_{si}|\mathbf{X}_d, \mathbf{H}_s) \leq \log\det\big(\lambda_i^2 \mathbf{X}_d^H \mathbf{X}_d + \frac{1}{\rho}\mathbf{I}_T\big), \tag{2.126}$$

$$= N_d \log\big(\lambda_i^2 + \frac{1}{\rho}\big) - (T - N_d)\log\rho, \tag{2.127}$$

and subsequently,

$$h(\mathbf{Y}_s''|\mathbf{X}_d, \mathbf{H}_s) \leq \sum_{i=1}^{N_d} \log(\lambda_i^2 + \frac{1}{\rho}) - N_d(T - N_d)\log\rho. \tag{2.128}$$

We now find a lower bound for $h(\mathbf{Y}_s''|\mathbf{H}_s)$. Similar to (2.102), we have

$$h(\mathbf{Y}_s''|\mathbf{H}_s) \geq h(\mathbf{X}_d) + \mathbb{E}[J_{X_s}], \tag{2.129}$$

where the expectation is with respect to $\mathbf{X}_s$, and

$$h(\mathbf{X}_d) = |\mathbb{F}(T, N_d)| = \prod_{i=T-N_d+1}^{T} \frac{2\pi^i}{(i-1)!}, \tag{2.130}$$

and

$$J_{X_s} = \prod_{i=1}^{N_d} \lambda^{2N_d} \det(\mathbf{X}_s)^{N_d}. \tag{2.131}$$

Finally, taking expectation over $\mathbf{H}_s$, we obtain

$$\begin{aligned}\mathbb{E}[I(\mathbf{X}_d; \mathbf{Y}_s|\mathbf{H}_s)] &\geq N_d(T - N_d)\log\rho + h(\mathbf{X}_d) + \mathbb{E}[J_{X_2}] - \mathbb{E}\big[\sum_{i=1}^{N_d} \log(\lambda_i^2 + \frac{1}{\rho})\big] \\ &= N_d(T - N_d)\log\rho + O(1). \end{aligned}\tag{2.132}$$

This completes the proof for Lemma 2.5.2.

Therefore, the transmitter is able to send $N_d(1 - N_d/T)$ DoF to the dynamic receiver, while ensuring the dynamic signal is decoded at the static receiver.

After decoding $\mathbf{X}_d$, the static receiver removes the interference:

$$\mathbf{Y}_s\mathbf{X}_d^H = \sqrt{\frac{T}{N_d N_s}}\mathbf{H}_s\mathbf{X}_s + \frac{1}{\sqrt{\rho}}\mathbf{W}_s', \tag{2.133}$$

where $\mathbf{W}_s' \in \mathcal{C}^{N_s \times N_d}$ is the equivalent noise whose entries are still i.i.d. $\mathcal{CN}(0,1)$. The equivalent channel for the static receiver is now a point-to-point MIMO channel. With Gaussian input $\mathbf{X}_s$, we have [26]

$$I(\mathbf{X}_s; \mathbf{Y}_s|\mathbf{H}_s) = N_d N_s \log\rho + O(1). \tag{2.134}$$

Normalizing $I(\mathbf{X}_s; \mathbf{Y}_s|\mathbf{H}_s)$ over T time-slots yields the achievable rate of the static receiver.

CHAPTER 3
COHERENT PRODUCT SUPERPOSITION FOR DOWNLINK MULTIUSER MIMO

3.1 Introduction

Due to varying mobilities, wireless network nodes often have unequal capability to acquire CSIR (channel state information at receiver). Downlink (broadcast) transmission to nodes with unequal CSIR is therefore a subject of practical interest.

It has been known that if all downlink users have full CSIR or none of them do, then orthogonal transmission (e.g. TDMA) achieves the optimal degrees of freedom (DoF) [13,14] when no CSIT is available. Chapter 2 ([30]) finds that a very different behavior emerges when one user has perfect CSIR and the other has none: in this case TDMA is highly suboptimal and a product superposition can achieve the optimal degrees of freedom (DoF). However, the analysis of Chapter 2 was limited to high-SNR, did not demonstrate optimality in all receiver antenna configurations, and more importantly, it required non-coherent Grassmannian signaling.

Most practical systems use pilots and employ channel estimation and coherent detection. Therefore in this chapter we extend the product superposition to coherent signaling with pilots. We show the DoF optimality of product superposition for more antenna configurations, and in addition show that it has excellent performance in low-SNR as well as high-SNR.

A downlink scenario with two users is considered in this chapter, where one user has a short coherence interval and is referred to as the *dynamic user*, and the other has a long coherence interval and is referred to as the *static user*. The main results of this chapter are as follows.

- We propose a new signaling structure that is a product of two matrices representing the signals of the static and dynamic user, respectively, where the data for both users are transmitted using coherent signaling.

- We propose two decoding methods. The first method performs no interference cancellation at the receiver. We show that under this method, at both high SNR and low SNR, the dynamic user experiences almost no degradation due to the transmission of the static user. Therefore in the sense of the cost to the other user, the static user's rate is added to the system "for free." Avoiding interference cancellation gives this method the advantage of simplicity.

- The second method further improves the static user's rate by allowing it to decode and remove the dynamic user's signal. This increases the effective SNR for the static user and provides further rate gain.

- We show that the product decompostion is DoF optimal when the dynamic user has either more, less or equal number of antennas as the static user. Previously [30] the DoF optimality was demonstrated only when the dynamic user had fewer or equal number of antennas compared with the static user.

- Finally we show how CSIT for the static user, whenever available, can be used to reduce the decoding complexity and further improve the rate for the static user.

The following notation is used throughout the chapter: for a matrix $\mathbf{A}$, the transpose is denoted with $\mathbf{A}^t$, the conjugate transpose with $\mathbf{A}^H$, the pesudo inverse with $\mathbf{A}^\dagger$ and the element in row i and column j with $[\mathbf{A}]_{ij}$. The $k \times k$ identity matrix is denoted with $\mathbf{I}_k$. The set of $n \times m$ complex matrices is denoted with $\mathcal{C}^{n\times m}$. We denote $\mathcal{CN}(0,1)$ as the circularly symmetric complex Gaussian distribution with zero mean and unit variance. For all variables the subscripts "s" and "d" stand as mnemonics for "static" and "dynamic", respectively, and subscripts "τ" and "δ" stand for "training" and "data."

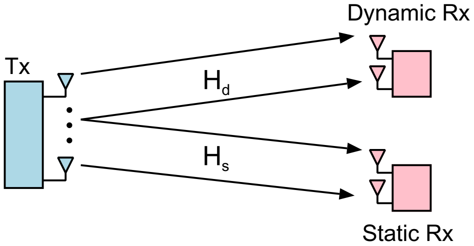

Figure 3.1. Channel model of Chapter 3.

3.2 System Model and Preliminaries

We consider an M-antenna base-station transmitting to two users, where the dynamic user has N_d antennas and the static user has N_s antennas. The channel coefficient matrices of the two users are $\mathbf{H}_d \in \mathcal{C}^{N_d \times M}$ and $\mathbf{H}_s \in \mathcal{C}^{N_s \times M}$, respectively. In this chapter we restrict our attention to $M = \max\{N_d, N_s\}$. The system operates under block-fading, where $\mathbf{H}_d$ and $\mathbf{H}_s$ remain constant for T_d and T_s symbols, respecitvely, and change independently across blocks. The coherence time T_d is small but T_s is large ($T_s \gg T_d$) due to different mobilities. The difference in coherence times means that the channel resources required by the static user to estimate its channel are negligible compared to the training requirements of the dynamic user. To reflect this in the model, it is assuemd that $\mathbf{H}_s$ is known by the static user (but unknown by the dynamic user, naturally), while $\mathbf{H}_d$ is not known *a priori* by either user.

Over T time-slots (symbols) the base-station sends $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_M]^t$ across M antennas, where $\mathbf{x}_i \in \mathcal{C}^{T \times 1}$ is the signal vector sent by the antenna i. The signal at the dynamic and static users is respectively

$$\begin{aligned} \mathbf{Y}_d &= \mathbf{H}_d \mathbf{X} + \mathbf{W}_d, \\ \mathbf{Y}_s &= \mathbf{H}_s \mathbf{X} + \mathbf{W}_s, \end{aligned} \tag{3.1}$$

where $\mathbf{W}_d \in \mathcal{C}^{N_d \times T}$ and $\mathbf{W}_s \in \mathcal{C}^{N_s \times T}$ are additive noise with i.i.d. entries $\mathcal{CN}(0, 1)$. Each row of $\mathbf{Y}_d \in \mathcal{C}^{N_d \times T}$ (or $\mathbf{Y}_s \in \mathcal{C}^{N_s \times T}$) corresponds to the received signal at an antenna of the dynamic user (or the static user) over T time-slots. The base-station is assumed to have an average power constraint ρ

$$\mathbb{E}\big[\sum_{i=1}^{M} \mathrm{tr}(\mathbf{x}_i\mathbf{x}_i^H)\big] = \rho\, T_d. \tag{3.2}$$

In this chapter, the base-station may know $\mathbf{H}_s$ (in Section 3.4) but does not know $\mathbf{H}_d$ due to its fast variation. The channels $\mathbf{H}_d$ and $\mathbf{H}_s$ have i.i.d. entries with the distribution $\mathcal{CN}(0,1)$. We assume $M = \max(N_d, N_s)$ and $T_d \geq 2N_d$ [20].

3.2.1 The Baseline Scheme

We start by establishing a baseline scheme and outlining its capacity for the purposes of comparison. In our system model, MIMO transmission schemes involving dirty paper coding, zero-forcing, or similar techniques [31–34] are not applicable since $\mathbf{H}_d$ varies too quickly for feedback to transmitter. Our baseline method uses orthogonal transmission, i.e., TDMA.

For the dynamic user, we consider the following near-optimal method. The base-station activates only N_d out of M antennas [20], sends an orthogonal pilot matrix $\mathbf{S}_\tau \in \mathcal{C}^{N_d\times N_d}$ during the first N_d time-slots, and then sends i.i.d. $\mathcal{CN}(0,1)$ data signal $\mathbf{S}_\delta \in \mathcal{C}^{N_d\times(T_d-N_d)}$ in the following $T_d - N_d$ time-slots [25], that is

$$\mathbf{X} = \Bigg[\sqrt{\frac{\rho_\tau}{N_d}}\,\mathbf{S}_\tau \;\sqrt{\frac{\rho_\delta}{N_d}}\,\mathbf{S}_\delta\Bigg] \tag{3.3}$$

where $\mathbf{S}_\tau\mathbf{S}_\tau^H = N_d\mathbf{I}$, and ρ_τ and ρ_δ are the average power used for training and data, respectively, and satisfy the power constraint in (3.2):

$$\rho_\tau N_d + \rho_\delta(T_d - N_d) \leq \rho T_d. \tag{3.4}$$

The dynamic user employs a linear minimum-mean-square-error (MMSE) estimation on the channel. The normalized channel estimate obtained in this orthogonal scheme is denoted $\overline{\mathbf{H}}_d \in \mathcal{C}^{N_d\times N_d}$. Under this condition, the rate attained by the dynamic user is [25]:

$$R_d \geq (1 - \frac{N_d}{T_d})\mathbb{E}\big[\log\det(\mathbf{I}_{N_d} + \frac{\rho_d}{N_d}\overline{\mathbf{H}}_d\overline{\mathbf{H}}_d^H)\big], \tag{3.5}$$

where ρ_d is the effective signal-to-noise ratio (SNR)

$$\rho_d = \frac{\rho_\delta\,\rho_\tau}{1 + \rho_\delta + \rho_\tau N_d}. \tag{3.6}$$

For the static user, the channel is known at the receiver, the base-station sends data directly using all M antennas. The rate achieved by the static user is [26]

$$R_s = \mathbb{E}\left[\log\det\left(\mathbf{I}_{N_s} + \frac{\rho}{N_s}\mathbf{H}_s\mathbf{H}_s^H\right)\right]. \tag{3.7}$$

Time-sharing between R_d and R_s yields the rate region

$$\mathcal{R}_{OT} = \big(tR_d,\, (1-t)R_s\big). \tag{3.8}$$

3.2.2 Overview of Product Superposition

In Chapter 2, a product superposition based on Grassmannian signaling was proposed and shown to achieve significant gain in DoF over orthogonal transmission. It has been shown that the method is DoF-optimal when $N_s \geq N_d$. In the so-called *Grassmannian-Ecludean superposition* [30], the base-station transmits

$$\mathbf{X} = \mathbf{X}_s\mathbf{X}_d \in \mathcal{C}^{N_s \times T_d} \tag{3.9}$$

over T_d time-slots, where $\mathbf{X}_d \in \mathcal{C}^{N_d \times T_d}$ and $\mathbf{X}_s \in \mathcal{C}^{N_s \times N_d}$ are the signals for the dynamic and static user, respectively. For the dynamic user, a Grassmannian (unitary) signal is used to construct $\mathbf{X}_d$, so that information is carried only in the subspace spanned by the rows of $\mathbf{X}_d$. As long as $\mathbf{X}_s$ is full rank, its multiplication does not create interference for the dynamic user, since $\mathbf{X}_s\mathbf{X}_d$ and $\mathbf{X}_d$ span the same row-space.

The static user decodes and peels off $\mathbf{X}_d$ from the received signal, then decodes $\mathbf{X}_s$, which carries information in the usual manner of space-time codes.

3.3 Pilot-Based Product Superposition

In this section, we develop a variation of product superposition that employs coherent signaling for both users. This is motivated by several factors, among them the popularity and prevalence of coherent signaling in the practice of wireless communications, as well as the known results in the point-to-point channel [20] showing that pilot-based transmission can perform almost as well as Grassmannian signaling. We show that a similar result holds in the mixed-mobility broadcast channel. The method proposed in this section uses single-user decoding (no interference cancellation).

3.3.1 Signaling Structure

Over T_d symbols (the coherence interval of the dynamic user) the base-station sends $\mathbf{X} \in \mathcal{C}^{N_s \times T_d}$ across N_s antennas:

$$\mathbf{X} = \mathbf{X}_s \mathbf{X}_d, \tag{3.10}$$

where $\mathbf{X}_s \in \mathcal{C}^{N_s \times N_d}$ is the data matrix for the static user and has i.i.d. $\mathcal{CN}(0,1)$ entries. The signal matrix $\mathbf{X}_d \in \mathcal{C}^{N_d \times T_d}$ is intended for the dynamic user and consists of the data matrix $\mathbf{X}_\delta \in \mathcal{C}^{N_s \times (T_d - N_s)}$ whose entries are i.i.d. $\mathcal{CN}(0,1)$ and the pilot matrix $\mathbf{X}_\tau \in \mathcal{C}^{N_s \times N_s}$ which is *unitary*, and is known to both static and dynamic users.

$$\mathbf{X}_d = \Big[\sqrt{c_\tau}\, \mathbf{X}_\tau \ \sqrt{c_\delta}\, \mathbf{X}_\delta \Big], \tag{3.11}$$

where the constant c_τ and c_δ satisfy the power constraint (3.2):

$$N_s N_d \big(c_\tau + (T_d - N_d) c_\delta \big) \leq \rho T_d. \tag{3.12}$$

Please make note of the normalization of pilot and data matrices in the product superposition: The pilot matrix is unitary, i.e., the entire pilot power is normalized, while the data matrix is normalized per time per antenna. This is only for convenience of mathematical expressions in the sequel; full generality is maintained via multiplicative constants c_δ and c_τ.

A sketch of the ideas involved in the decoding at the dynamic and static users is as follows. The signal received at the dynamic user is

$$\mathbf{Y}_d = \mathbf{H}_d \mathbf{X}_s \Big[\sqrt{c_\tau} \mathbf{X}_\tau \ \sqrt{c_\delta} \mathbf{X}_\delta \Big] + \mathbf{W}_d \tag{3.13}$$

where $\mathbf{W}_d$ is the additive noise. The dynamic user uses the pilot matrix to estimate the equivalent channel $\mathbf{H}_d \mathbf{X}_s$, and then decodes $\mathbf{X}_\delta$ based on the channel estimate.

For the static user, the signal received during the first N_d time-slots is

$$\mathbf{Y}_{s1} = \sqrt{c_\tau}\, \mathbf{H}_s \mathbf{X}_s \mathbf{X}_\tau + \mathbf{W}_{s1} \tag{3.14}$$

where $\mathbf{W}_{s1}$ is the additive noise at the static user during the first N_d samples. The static user multiplies its received signal by $\mathbf{X}_\tau^H$ from the right and then recovers the signal $\mathbf{X}_s$.

Remark 3.3.1 *Each of the dynamic user's codewords includes pilots because it needs frequent channel estimates. No pilots are included in the individual codewords of the static user because it only needs infrequent channel estimate updates. In practice static user's channel training occurs at much longer intervals outside the proposed signaling structure.*

3.3.2 Main Result

Theorem 3.3.1 *Consider an M-antenna base-station, a dynamic user with N_d-antennas and coherence time T_d, and a static user with N_s-antennas and coherence time $T_s \gg T_d$. Assuming the dynamic user does not know its channel $\mathbf{H}_d$ but the static user knows its channel $\mathbf{H}_s$, the pilot-based product superposition achieves the rates*

$$R_d = (1 - \frac{N_d}{T_d})\mathbb{E}\Big[\log\det\Big(\mathbf{I}_{N_d} + \frac{\rho_d}{N_d}\overline{\mathbf{H}}_d\overline{\mathbf{H}}_d^H\Big)\Big], \tag{3.15}$$

$$R_s = \frac{N_d}{T_d}\mathbb{E}\Big[\log\det\Big(\mathbf{I}_{N_s} + \frac{\rho_s}{N_s}\mathbf{H}_s\mathbf{H}_s^H\Big)\Big], \tag{3.16}$$

where $\overline{\mathbf{H}}_d$ is the normalized *MMSE channel estimate of the equivalent dynamic channel $\mathbf{H}_d\mathbf{X}_s$, and ρ_d and ρ_s are the effective SNRs:*

$$\rho_d = \frac{c_\tau c_\delta N_d N_s^2}{1 + c_\tau N_s + c_\delta N_d N_s}, \tag{3.17}$$

$$\rho_s = c_\tau N_s. \tag{3.18}$$

Proof See Section 3.6.1.

For the static user, the effective SNR $\rho_s = c_\tau$ increases linearly with the power used in the training of the dynamic user. This is because the static user decodes based on the signal received during the training phase of the dynamic user.

For the dynamic user, the effective SNR ρ_d is unaffected by superimposing $\mathbf{X}_s$ on $\mathbf{X}_d$. To see this, compare (3.4) with (3.12) to arrive at $\rho_\tau = c_\tau N_s$ and $\rho_\delta = c_\delta N_d N_s$, therefore the two SNRs are equal to

$$\rho_d = \frac{c_\tau c_\delta N_d N_s^2}{1 + c_\tau N_s + c_\delta N_d N_s}. \tag{3.19}$$

Intuitively, the rate available to the dynamic user via orthogonal transmission (Eq. (3.5)) and via superposition (Eq. (3.15)) will be very similar: the normalized channel estimate $\overline{\mathbf{H}}_d$ in both cases has uncorrelated entries with zero mean and unit variance.[1] Thus the product superposition achieves the static user's rate "for free" in the sense that the rate for the dynamic user is approximately the same as in the single-user scenario. In the following, we discuss this phenomenon at low and high SNR.

Low-SNR Regime

We have $\rho_d, \rho_s \ll 1$. Let the eigenvalues of $\overline{\mathbf{H}}_d \overline{\mathbf{H}}_d^H$ be denoted $\bar{\lambda}_{1i}^2$, $i = 1, \ldots, N_d$. Using (3.15) and a Taylor expansion of the log function at low SNR, the achievable rate for the dynamic user is approximately:

$$R_d \approx (1 - \frac{N_d}{T_d}) \frac{\rho_d}{N_d} \, \mathbb{E}\Big[\sum_{i=1}^{N_d} \bar{\lambda}_{1i}^2 \Big] \tag{3.20}$$

$$= (1 - \frac{N_d}{T_d}) \frac{\rho_d}{N_d} \, \mathrm{tr}\big(\mathbb{E}[\overline{\mathbf{H}}_d \overline{\mathbf{H}}_d^H] \big) \tag{3.21}$$

$$= (1 - \frac{N_d}{T_d}) N_d \, \rho_d. \tag{3.22}$$

Similarly, from (3.5), the baseline method achieves the rate

$$(1 - \frac{N_d}{T_d}) N_d \, \rho_d. \tag{3.23}$$

[1] The dynamic channel estimates in the orthogonal and superposition transmissions have the same mean and variance but are not identically distributed, becuase in the orthogonal case, $\overline{\mathbf{H}}_d$ is an estimate of $\mathbf{H}_d$, a Gaussian matrix, while in the superposition case it is an estimate of $\mathbf{H}_d \mathbf{X}_s$, the product of two Gaussian matrices. Therefore the expectations in Eq. (3.5) and (3.15) may produce slightly different results.

Thus, the dynamic user attains the same rate as it would in the absence of the other user and its interference, i.e., a single-user rate. At low SNR, one cannot exceed this performance.

The rate available to the static user at low-SNR is obtained via (3.16), as follows:

$$R_s \approx \frac{\rho_s}{T_d}\operatorname{tr}\Big(\mathbb{E}[\mathbf{H}_s\mathbf{H}_s^H]\Big) \tag{3.24}$$

$$= \frac{N_s^2\,\rho_s}{T_d}. \tag{3.25}$$

High-SNR Regime

We have $\rho_d, \rho_s \gg 1$, therefore from (3.15) the achievable rate for the dynamic user is

$$R_d \approx (1-\frac{N_d}{T_d})\Big(N_d \log\frac{\rho_d}{N_d} + \mathbb{E}\big[\sum_{i=1}^{N_d}\log\bar{\lambda}_{1i}^2\big]\Big). \tag{3.26}$$

The dynamic user attains $N_d(1-N_d/T_d)$ degrees of freedom, which is the maximum DoF even in the absence of the static user [20]. Superimposing $\mathbf{X}_s$ only affects the distribution of eigenvalues $\bar{\lambda}_{1i}^2$, whose impact is negligible at high-SNR.

For the static user, let the eigenvalues of $\mathbf{H}_s\mathbf{H}_s^H$ be denoted λ_{2i}^2, $i = 1, \ldots, N_s$. From (3.16), we have

$$R_s \approx \frac{N_d}{T_d}\Big(N_s \log\frac{\rho_s}{N_s} + \mathbb{E}\big[\sum_{i=1}^{N_s}\log\lambda_{2i}^2\big]\Big), \tag{3.27}$$

which implies that the static user achieves N_dN_s/T_d degrees of freedom. Thus, the pilot-based product superposition achieves the optimal DoF obtained in [30] for $N_d \le N_s$, and for $N_d > N_s$ meets the coherent upper bound.

3.3.3 Power Allocation

The effective SNRs of the dynamic and static users depend on c_τ and c_δ. We focus on c_τ and c_δ that maximize R_d (equivalently ρ_d) in a manner similar to [25]. From (3.79) and (3.72),

$$\rho_d = \frac{c_\tau c_\delta N_d N_s^2}{1 + c_\tau N_s + c_\delta N_d N_s}. \tag{3.28}$$

From (3.12), we have $c_\tau = \rho T_d/(N_d N_s) - c_\delta(T_d - N_d)$. Substitue c_τ into (3.28):

$$\rho_d = \frac{N_d N_s(T_d - N_d)}{T_d - 2N_d} \cdot \frac{c_\delta(a - c_\delta)}{-c_\delta + b}, \tag{3.29}$$

where

$$a = \frac{\rho T_d}{N_d N_s(T_d - N_d)}, \tag{3.30}$$

$$b = \frac{N_d + \rho T_d}{N_d N_s(T_d - 2N_d)}. \tag{3.31}$$

Noting that $0 \le c_\delta \le a$, we obtain the value of c_δ that maximizes R_d:

$$c_\delta^* = b - \sqrt{b^2 - ab}, \tag{3.32}$$

which corresponds to

$$\rho_d^* = \frac{N_d N_s(T_d - N_d)}{T_d - 2N_d}\big(2b - a - 2\sqrt{b^2 - ab}\big), \tag{3.33}$$

$$\rho_s^* = \frac{\rho T_d}{N_d} - N_s(T_d - N_d)(b - \sqrt{b^2 - ab}). \tag{3.34}$$

In the low-SNR regime where $\rho \ll 1$ we have

$$\rho_d^* \approx \frac{\rho^2 T_d^2}{4(T_d - N_d)} \tag{3.35}$$

$$\rho_s^* \approx \frac{\rho T_d}{2N_d}. \tag{3.36}$$

This indicates that the static user has a much larger effective SNR, i.e., $\rho_d^* = o(\rho_s^*)$. In this case, from (3.22) and (3.25), the achievable rate is

$$R_d \ge \frac{T_d N_d}{4}\rho^2, \tag{3.37}$$

$$R_s \approx \frac{N_s}{2}\rho. \tag{3.38}$$

In the high-SNR regime where $\rho \gg 1$ we have

$$\rho_d^* \approx \frac{\rho\, T_d}{(\sqrt{T_d - N_d} - \sqrt{N_d})^2}, \tag{3.39}$$

$$\rho_s^* \approx \frac{\rho T_d(\sqrt{T_d/N_d - 1} - 1)}{T_d - 2N_d}. \tag{3.40}$$

Both static and dynamic users attain SNR that increases linearly with ρ. When $T_d \gg N_d$, for the static user, $\rho_s^* \approx \rho\sqrt{T_d/N_d} \gg \rho_d^*$. For the dynamic user, we have $\rho_d^* \approx \rho$, which is the same SNR as if the dynamic user had perfect CSI; this is not suprising since the power used for training is negligible when the channel is very steady.

Remark 3.3.2 *In the MIMO broadcast channel, conventional transmission schemes essentially divide the power between users. In the proposed product superposition the transmit power works for both users simultaneously instead of being divided between them. The training power used for the dynamic user also carries the static user's data. In this way, significant gains over TDMA are achieved, which is contrary to the conventional methods that at low-SNR produce little or no gain relative to TDMA.*

Remark 3.3.3 *Note that the developments in this section make no assumption about the relative number of antennas at the dynamic and static receivers, yet in Equations* (3.15) *and* (3.16) *the degrees of freedom (prelog factors) meet upper the bound [30] when $N_d \leq N_s$ as well as the coherent upper bound when $N_d > N_s$. Thus, the DoF optimality of the product superposition, which was shown in [30] only for $N_d \leq N_s$, is now extended for all dynamic/static user antenna configurations.*

3.4 Improving Rates by Interference Cancellation and Partial CSIT

3.4.1 Interference Decoding and Cancellation

We focus on the case of $N_s \geq N_d$. Previously, the static user could only use the portion of power that was shared with the dynamic user pilot (but not the dynamic user data). However, if the static user has no fewer antennas than the dynamic user, in principle it may decode the dynamic user's data and then harvest the entire power of the signal.

The received signal at the static user is

$$\mathbf{Y}_s = \mathbf{H}_s\mathbf{X}_s[\sqrt{c_\tau}\,\mathbf{X}_\tau \;\sqrt{c_\delta}\,\mathbf{X}_\delta] + \mathbf{W}_s \tag{3.41}$$

where $\mathbf{Y}_s \in \mathcal{C}^{N_s \times T_d}$. The static user first estimates the product $\mathbf{H}_s\mathbf{X}_s \in \mathcal{C}^{N_s \times N_d}$ by using the pilot $\mathbf{X}_\tau$ sent during the first N_d time-slots, and then it may deocde $\mathbf{X}_\delta$. Given the entire $\mathbf{X}_d$, ML decoding can be used to decode the static user's signal. Assuming the codeword used by the dynamic user is sufficiently long, the rate gain produced by the interference decoding is characterized by the following theorem.

Theorem 3.4.1 **($N_s \geq N_d$)** *With interference decoding and cancellation, the pilot-based product superposition achieves the following rate for the static user*

$$R_s = \frac{N_d}{T_d}\mathbb{E}\left[\log\det\left(\mathbf{I}_{N_s} + \frac{\rho_s}{N_s}\mathbf{H}_s\mathbf{H}_s^H\right)\right], \tag{3.42}$$

where the effective SNR is

$$\rho_s = \frac{N_s}{\mathbb{E}[\lambda_i^{-2}]} \tag{3.43}$$

with λ_i^2 being any of the unorderred eigenvalues of $\mathbf{X}_d\mathbf{X}_d^H$.

Proof See Section 3.6.2.

Compared with Theorem 3.3.1, the SNR for the static user is improved by using the entire $\mathbf{X}_d$. To see this, we decompose $\mathbf{X}_\delta = \mathbf{U}_\delta \operatorname{diag}(\gamma_1, \cdots, \gamma_{N_d})\, \mathbf{V}_\delta^H$, and obtain

$$\mathbf{X}_d\mathbf{X}_d^H = c_\tau \mathbf{I}_{N_d} + c_\delta\, \mathbf{U}_\delta \operatorname{diag}(\gamma_1^2, \cdots, \gamma_{N_d}^2)\, \mathbf{U}_\delta^H \tag{3.44}$$

$$= \operatorname{diag}(c_\tau + c_\delta\gamma_1^2, \cdots, c_\tau + c_\delta\gamma_{N_d}^2). \tag{3.45}$$

Therefore, $\lambda_i^2 = c_\tau + c_\delta\, \gamma_i^2$, for $i = 1, \ldots, N_d$, and

$$\rho_s = \frac{N_s}{\mathbb{E}[(c_\tau + c_\delta\, \gamma_1^2)^{-1}]}. \tag{3.46}$$

which is greater than the effective power available to the previous scheme (compare with Eq. (3.18)). So knowing the dynamic user's data always produces a power gain. A closed form expression for ρ_s is not tractable so we find lower and upper bounds.

$$\mathbb{E}[\gamma_i^2] = \frac{1}{N_d}\mathbb{E}[\sum_{i=1}^{N_d}\gamma_i^2] = \frac{1}{N_d}\mathbb{E}[\operatorname{tr}(\mathbf{X}_\delta\mathbf{X}_\delta^H)] \tag{3.47}$$

$$= T_d - N_d. \tag{3.48}$$

Due to convexity of the function $1/x$, we have

$$\mathbb{E}[(c_\tau + c_\delta\, \gamma_i^2)^{-1}] \geq \left(c_\tau + c_\delta \mathbb{E}[\gamma_i^2]\right)^{-1} \tag{3.49}$$

$$= \frac{1}{c_\tau + c_\delta (T_d - N_d)}, \tag{3.50}$$

and thus

$$\rho_s \leq c_\tau N_s + c_\delta N_s (T_d - N_d). \tag{3.51}$$

Recall that $\mathbf{X}_s$ has N_d columns, so the total power used for the static user is $\rho_s N_d$. Therefore, from (3.12), the upper bound corresponds to the case where the static user collects all the transmit power.

For the lower bound, we use the fact that the arithmetic mean is no less than the harmonic mean, and obtain

$$\mathbb{E}[(c_\tau + c_\delta\, \gamma_i^2)^{-1}] = \mathbb{E}\left[\frac{1}{1/c_\tau^{-1} + 1/(c_\delta \gamma_i^2)^{-1}}\right] \tag{3.52}$$

$$\leq \frac{1}{4}\left(c_\tau^{-1} + c_\delta^{-1}\mathbb{E}[\gamma_i^{-2}]\right) \tag{3.53}$$

$$= \frac{1}{4}\left(c_\tau^{-1} + c_\delta^{-1}(T_d - N_d - 1)^{-1}\right), \tag{3.54}$$

where (3.54) uses the fact that γ_i^{-2} has inverse Gamma distribution with mean $1/(T_d - N_d - 1)$.

3.4.2 Partial CSIT

In each of the methods mentioned earlier in this section, the static user operates under an equivalent single-user channel, by inverting either the pilot component or all components of the dynamic user's signal. Thus, any benefits that can be realized in the single-user MIMO can also be available to the static user, including the benefits arising from CSIT. For example, water-filling can be applied to allocate power across multiple eigen-modes of the static user. CSI can also simplify decoding at the static user. To see this, using singular value decomposition (SVD), $\mathbf{H}_s = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H$, where $\mathbf{\Sigma}_s = \text{diag}(\lambda_1, \cdots, \lambda_{N_s})$. Then, the base-station sends

$$\mathbf{X} = \mathbf{V}\mathbf{X}_s\, [\sqrt{c_\tau}\mathbf{X}_\tau\ \sqrt{c_\delta}\mathbf{X}_\delta]. \tag{3.55}$$

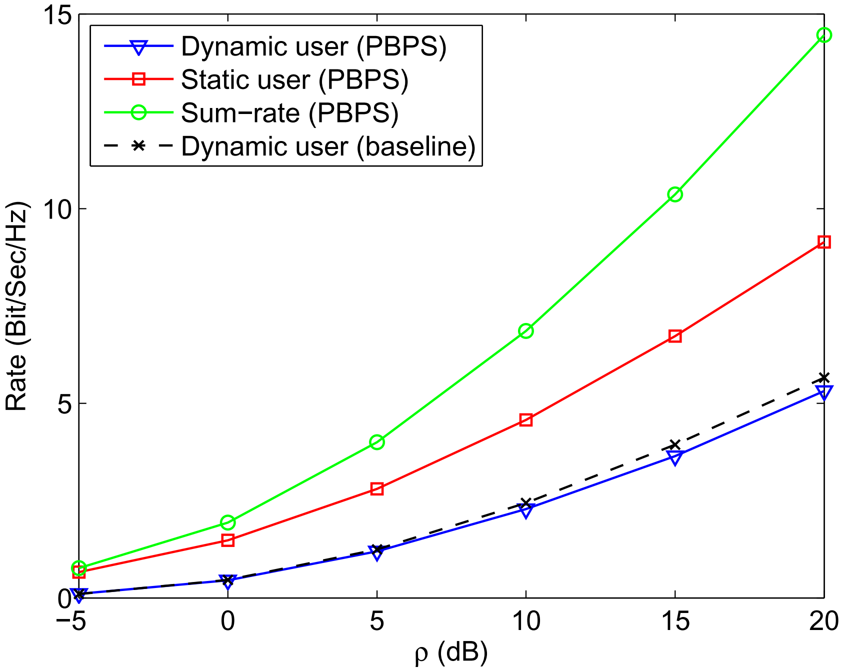

Figure 3.2. Pilot-based product superposition (PBPS), $N_d = 2$, $N_s = M = 4$ and $T_d = 5$.

Since $\mathbf{V}$ is unitary, the entries of $\mathbf{V}\mathbf{X}_s$ remain i.i.d. $\mathcal{CN}(0, 1)$, and therefore, the performance of the dynamic user is unaffected by precoding with $\mathbf{V}$. Without interference decoding, the static user forms the equivalent diagonal channel

$$\mathbf{U}^H \mathbf{Y}_\tau \mathbf{X}_\tau = \sqrt{c_\tau}\, \mathbf{X}_s + \mathbf{W}'_\tau, \tag{3.56}$$

where $\mathbf{W}'_\tau$ is the noise with i.i.d. $\mathcal{CN}(0, 1)$ distribution.

3.5 Numerical Results

Unless specified otherwise, a power allocation is assumed (c_τ and c_δ) that maximizes the rate for the dynamic user.

Figure 3.2 illusrates the rate for dynamic and staic users in the pilot-based product superposition, as shown in Theorem 3.3.1. We consider $N_d = 2$, $N_s = M = 4$ and $T_d = 5$. Both the baseline method and proposed methods optimize the rate for the dynamic user.

In this case, the baseline method cannot provide any rate for the static user. In addition to near-optimal rate for the dynamic user, the proposed method significant rate for the static user. The separation from optimality is negligible in the low-SNR regime, and in the high-SNR regime the rate of the dynamic user has the optimal degrees of freedom (SNR slope).

Thus the proposed method achieves the static user's rate almost "for free" in terms of the penalty to the other user.

Figure 3.3 shows the impact of the available antenna of the static user. Here, $\rho = 10$ dB, $N_d = 2$, $M = N_s$ and $T_d = 5$. The static user's rate (thus the sum-rate) increases linearly with N_s, because the degrees of freedom is $N_d N_s / T_d$, as indicated by Theorem 3.3.1. The gap of the dynamic user's rate under the proposed method and the baseline method vanishes as N_s increases. Intuitively, the rate difference is because of the Jensen's loss: in the proposed method the equivalent channel is the product matrix $\mathbf{H}_d \mathbf{X}_s$ and is "more spread" than the channel in the baseline method. As N_s increases, $\mathbf{X}_s$ becomes "more unitary" ($\mathbf{X}_s \mathbf{X}_s^H / N_s \to \mathbf{I}_{N_d}$) and thus less impact on the distribution of $\mathbf{H}_d$.

The power allocated in the training and data period is $c_\tau N_d N_s$ and $c_\delta N_d N_s (T_d - N_d)$, respectively. The impact of different power allocation is shown by Figure 3.4, where $\rho = 5$ dB, $N_d = 2$, $N_s = M = 4$ and $T_d = 5$. In Figure 3.4, the larger the ratio c_τ / c_δ is, the more power is used for training. The optimal ratio that maximizes the rate for the dynamic user can be derived from (3.32) and is similar to the result in [25]. The static user's rate increases logrimathtically with c_τ / c_δ, because without interference decoding, the effective power received by the static user is proportional to c_τ.

In Figure 3.5, we investigate the impact of the channel coherence time of the dynamic user. Here, $\rho = 5$ dB, $N_d = 2$, and $N_s = M = 4$. As T_d increases, the rate for the dynamic user improves, since the portion of time-slots (overhead) used for training is reduced. In contrast, the rate for the static user decreases with T_d, because the static user transmits new signal matrix over T_d period. Intuitively, as T_d increases, the dynamic user's channel

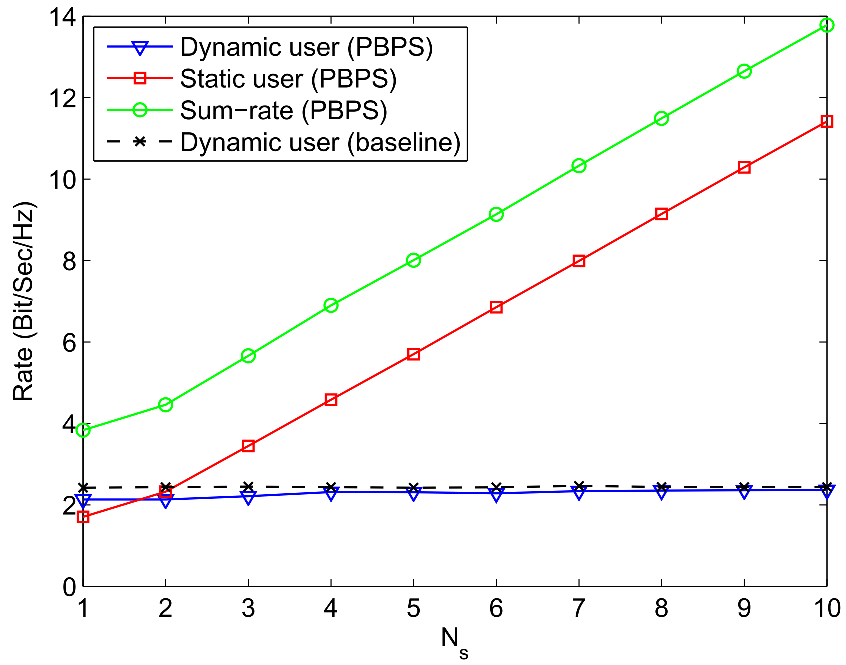

Figure 3.3. Impact of the static user's antennas : $\rho = 10$ dB, $N_d = 2$, $M = N_s$ and $T_d = 5$.

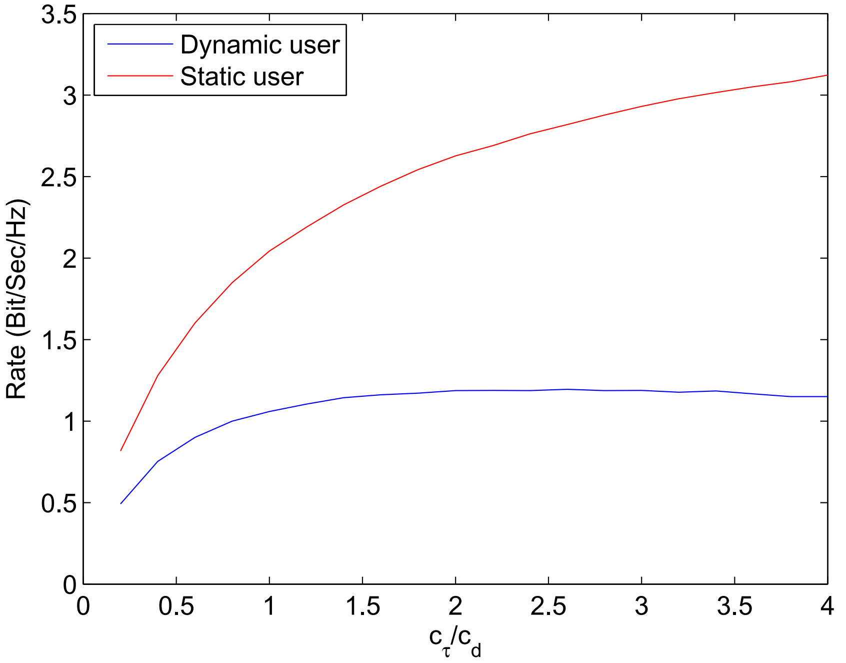

Figure 3.4. Impact of power allocation: $\rho = 5$ dB, $N_d = 2$, $N_s = M = 4$ and $T_d = 5$.

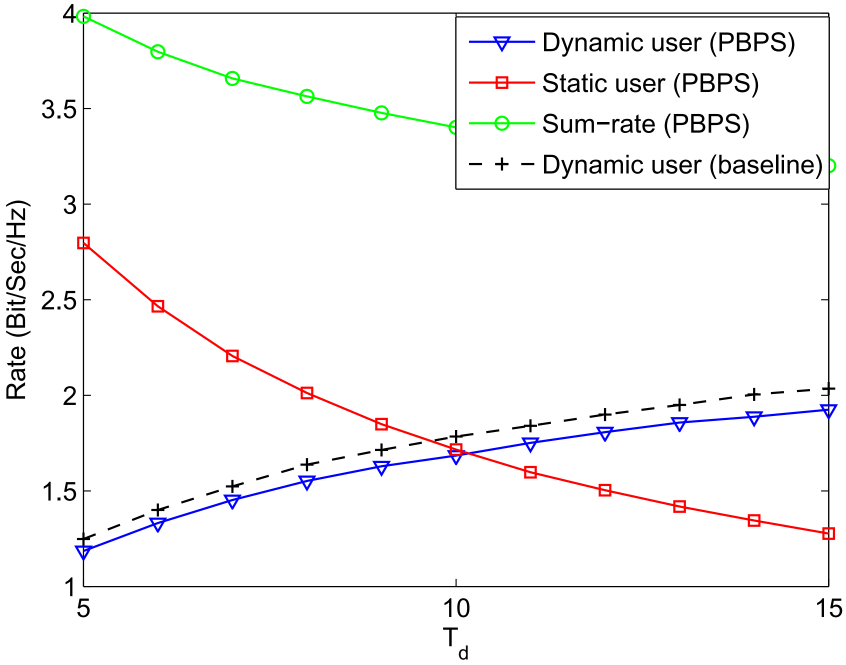

Figure 3.5. Impact of channel coherence time: $\rho = 5$ dB, $N_d = 2$, and $N_s = M = 4$.

becomes “more static”, and therefore, the opportunity to explore its “insensitivity” to the channel is reduced.

Finally, in Figure 3.6, we show the gain of interference decoding in the pilot-based product superposition, where $N_d = 2$, $N_s = M = 4$ and $T_d = 5$. By decoding the dynamic signal and harvesting the power carried by the entire data, the static rate is improved around 10%.

3.6 Proof of Theorem

3.6.1 Proof of Theorem 3.3.1

Rate of the Static User

During the first N_d time-slots, the static user receives

$$\mathbf{Y}_{s1} = \sqrt{c_\tau}\,\mathbf{H}_s\mathbf{X}_s\mathbf{X}_\tau + \mathbf{W}_{s1}. \tag{3.57}$$

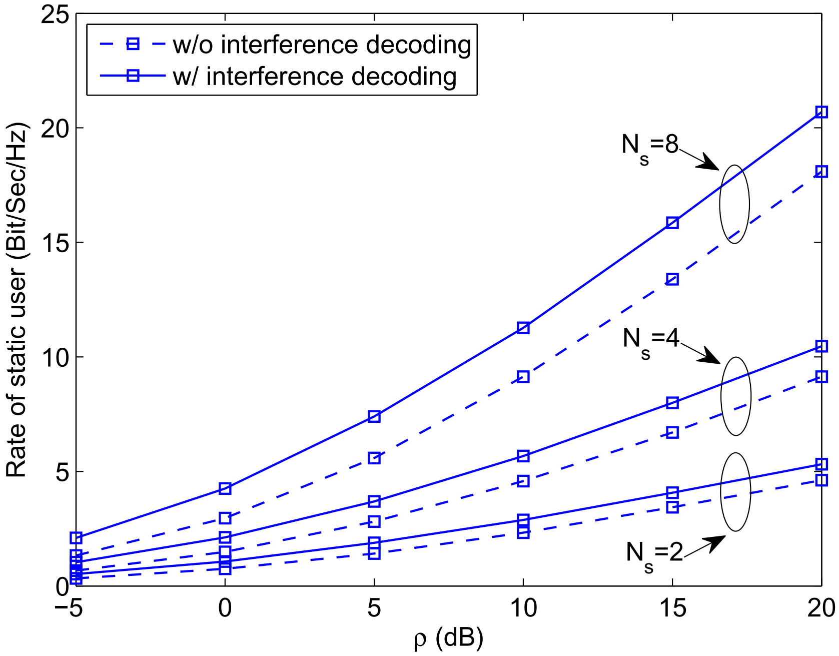

Figure 3.6. Static user's rate with interference decoding: $N_d = 2$, $N_s = M = 4$ and $T_d = 5$.

Because the static user knows $\mathbf{X}_\tau$, it removes the impact of $\mathbf{X}_\tau$ from $\mathbf{Y}_{2\tau}$:

$$\mathbf{Y}'_{s1} = \mathbf{Y}_{s1}\mathbf{X}_\tau^H \tag{3.58}$$

$$= \sqrt{c_\tau}\,\mathbf{H}_s\mathbf{X}_s + \mathbf{W}'_{s1} \tag{3.59}$$

where $\mathbf{Y}_{s1} \in \mathcal{C}^{N_s \times N_d}$ and $\mathbf{W}'_{s1}$ is the equivalent noise whose entries remain i.i.d. $\mathcal{CN}(0,1)$. Therefore, the channel seen by the static user becomes a point-to-point MIMO channel. Let $\mathbf{y}'_{si}$ and $\mathbf{x}_{si}$ be the column i of $\mathbf{Y}'_{s1}$ and $\mathbf{X}_s$, respectively. The mutual information

$$I(\mathbf{Y}_{s1}; \mathbf{X}_s) = \sum_{i=1}^{N_d} I(\mathbf{y}'_{si}; \mathbf{x}_{si}) \tag{3.60}$$

$$= N_d \log\det\left(\mathbf{I}_{N_s} + c_\tau\,\mathbf{H}_s\mathbf{H}_s^H\right), \tag{3.61}$$

which implies that the effective SNR for the static user is

$$\rho_s = c_\tau. \tag{3.62}$$

In the following $T_d - N_d$ time-slots, the static user disregards the received signal. The average rate achieved by the static user is

$$R_s = \frac{N_d}{T_d}\,\mathbb{E}\left[\log\det\left(\mathbf{I}_{N_s} + \rho_s\,\mathbf{H}_s\mathbf{H}_s^H\right)\right], \tag{3.63}$$

where the expectation is over the channel realizations of $\mathbf{H}_s$.

Rate of the Dynamic User

The dynamic user first estimates the equivalent channel and then decodes its data. During the first N_d time-slots, the dynamic user receives the pilot signal

$$\mathbf{Y}_\tau = \sqrt{c_\tau}\,\mathbf{H}_d\mathbf{X}_s\mathbf{X}_\tau + \mathbf{W}_\tau \tag{3.64}$$

$$= \sqrt{c_\tau N_s}\,\widetilde{\mathbf{H}}_d\mathbf{X}_\tau + \mathbf{W}_\tau, \tag{3.65}$$

where $\widetilde{\mathbf{H}}_d \in \mathcal{C}^{N_d\times N_d}$ is the equivalent channel of the dynamic user

$$\widetilde{\mathbf{H}}_d \triangleq \frac{1}{\sqrt{N_s}}\,\mathbf{H}_d\mathbf{X}_s \tag{3.66}$$

Let $\tilde{h}_{ij} = [\widetilde{\mathbf{H}}_d]_{ij}$, then we have $\mathbb{E}[\tilde{h}_{ij}] = 0$ and

$$\mathbb{E}[\tilde{h}_{ij}\,\tilde{h}_{pq}^H] = \begin{cases} 1, & \text{if } (i,j) = (p,q) \\ 0, & \text{else} \end{cases}, \tag{3.67}$$

i.e., the entries of $\widetilde{\mathbf{H}}_d$ are uncorrelated and have zero-mean and unit variance.

The dynamic user estimates $\widetilde{\mathbf{H}}_d$ by the MMSE. Let

$$C_{YY} = (1 + c_\tau N_s)\mathbf{I}_{N_d}, \quad C_{YH} = \sqrt{c_\tau N_s}\,\mathbf{X}_\tau^H, \tag{3.68}$$

we have

$$\widehat{\mathbf{H}}_d = \mathbf{Y}_\tau C_{YY}^{-1} C_{YH} \tag{3.69}$$

$$= \frac{\sqrt{c_\tau N_s}}{1 + c_\tau N_s}\left(\sqrt{c_\tau N_s}\,\widetilde{\mathbf{H}}_d + \mathbf{W}_\tau\mathbf{X}_\tau^H\right) \tag{3.70}$$

Because $\mathbf{W}_\tau$ has i.i.d. $\mathcal{CN}(0,1)$ entries, the noise matrix $\mathbf{W}_\tau \mathbf{X}_\tau^H$ also has i.i.d. $\mathcal{CN}(0,1)$ entries. Define $\hat{h}_{1ij} = [\widehat{\mathbf{H}}_d]_{ij}$. Then, we have $\mathbb{E}[\hat{h}_{1ij}] = 0$ and

$$\mathbb{E}[\hat{h}_{ij}\hat{h}_{pq}^H] = \begin{cases} \delta_1^2, & \text{if } (i,j) = (p,q) \\ 0, & \text{else} \end{cases}, \tag{3.71}$$

where

$$\delta_1^2 = \frac{c_\tau N_s}{1 + c_\tau N_s}. \tag{3.72}$$

In other words, the estimate of the equivalent channel has uncorrelated elements with zero-mean and variance δ_1^2.

During the remaining $T_d - N_d$ time-slots, the dynamic user regards the channel estimate $\widehat{\mathbf{H}}_d$ as the true channel and decodes the data signal. At the time-slot i, $N_d < i \le T_d$, the dynamic user receives

$$\mathbf{y}_{di} = \sqrt{c_\delta N_s}\,\widehat{\mathbf{H}}_d \mathbf{x}_{di} + \underbrace{\sqrt{c_\delta N_s}\,\widetilde{\mathbf{H}}_e \mathbf{x}_{1di} + \mathbf{w}_{di}}_{\mathbf{w}'_{di}}, \tag{3.73}$$

where $\widetilde{\mathbf{H}}_e = \widetilde{\mathbf{H}}_d - \widehat{\mathbf{H}}_d$ is the estimation error for $\widetilde{\mathbf{H}}_d$, and $\mathbf{w}'_{di}$ is the equivalent noise that has zero mean and autocorrelation

$$\mathbf{R}_{w_d} = c_\delta N_s\, \mathbb{E}\big[\widetilde{\mathbf{H}}_e \widetilde{\mathbf{H}}_e^H\big] + \mathbf{I}_{N_d} \tag{3.74}$$

$$= \big(1 + \frac{c_\delta N_d N_s}{1 + c_\tau N_s}\big)\mathbf{I}_{N_d}. \tag{3.75}$$

Using the argument that Gaussian distribution maximizes the differential entroy with given second moments [17], the mutual information is lower-bounded as

$$I(\mathbf{y}_{di}; \mathbf{x}_{di} | \widehat{\mathbf{H}}_d) \ge \log\det\left(\mathbf{I}_{N_d} + \frac{c_\delta N_s\, \widehat{\mathbf{H}}_d \widehat{\mathbf{H}}_d^H}{1 + c_\delta N_d N_s/(1 + c_\tau N_s)}\right) \tag{3.76}$$

$$= \log\det\left(\mathbf{I}_{N_d} + \frac{c_\delta \delta_1^2 N_s\, \overline{\mathbf{H}}_d \overline{\mathbf{H}}_d^H}{1 + c_\delta N_d N_s/(1 + c_\tau N_s)}\right), \tag{3.77}$$

where $\overline{\mathbf{H}}_d$ is the normorlized channel whose elements have unit variance

$$\overline{\mathbf{H}}_d = \frac{1}{\delta_1}\widehat{\mathbf{H}}_d. \tag{3.78}$$

From (3.77), the effective SNR for the dynamic user is

$$\rho_d = \frac{c_\delta \delta_1^2 N_s}{1 + c_\delta N_d N_s/(1 + c_\tau N_s)}. \tag{3.79}$$

The average rate that the dynamic user achieves is

$$R_d \geq (1 - \frac{N_d}{T_d})\mathbb{E}\big[\log\det(\mathbf{I}_{N_d} + \rho_d \overline{\mathbf{H}}_d \overline{\mathbf{H}}_d^H)\big], \tag{3.80}$$

where the expectation is over the dynamic user's channel realizations.

3.6.2 Proof of Theorem 3.4.1

We first show that if the codeword used by the dynamic user is sufficiently long, the static user always decodes the dynamic user's signal. Then, we find the rate for the static user given the knowledge of the dynamic user's signal.

During the first N_d time-slots, the static user receives

$$\mathbf{Y}_{s1} = \sqrt{c_\tau N_s}\,\widetilde{\mathbf{H}}_s \mathbf{X}_\tau + \mathbf{W}_{s1}, \tag{3.81}$$

where $\widetilde{\mathbf{H}}_s \triangleq \mathbf{H}_s \mathbf{X}_s/\sqrt{N_s}$ is the composite channel of the static user. Define $\tilde{h}_{ij} = [\widetilde{\mathbf{H}}_s]_{ij}$. Then, we have $\mathbb{E}[\tilde{h}_{ij}] = 0$ and

$$\mathbb{E}[\tilde{h}_{ij}\tilde{h}_{pq}^H] = \begin{cases} 1, & \text{if } (i,j) = (p,q) \\ 0, & \text{else} \end{cases}, \tag{3.82}$$

i.e., the entries of $\widetilde{\mathbf{H}}_s$ are uncorrelated and have zero-mean and unit variance. For simplicity, we do not exploit the knowledge of $\mathbf{H}_s$, i.e., viewing $\widetilde{\mathbf{H}}_s$ as a product of two random matrices, and use the MMSE estimation similar to Section 3.3:

$$\widehat{\mathbf{H}}_s = \frac{\sqrt{c_\tau N_s}}{1 + c_\tau N_s}\mathbf{Y}_{s1}\mathbf{X}_\tau^H. \tag{3.83}$$

Define $\bar{h}_{ij} = [\widehat{\mathbf{H}}_s]_{ij}$. Then, we have $\mathbb{E}[\bar{h}_{ij}] = 0$ and

$$\mathbb{E}[\bar{h}_{ij}\bar{h}_{pq}^H] = \begin{cases} \delta_{\hat{H}_s}^2, & \text{if } (i,j) = (p,q) \\ 0, & \text{else} \end{cases}, \tag{3.84}$$

where $\delta^2_{\hat{H}_s} = \delta_1^2$. Based on $\widehat{\mathbf{H}}_d$ the static user decodes the dynamic signal. At time-slot $i \in \{N_d, \ldots, T_d\}$ the static user receives

$$\mathbf{y}_{si} = \sqrt{c_\delta N_s}\, \widehat{\mathbf{H}}_s \mathbf{x}_{di} + \underbrace{\sqrt{c_\delta N_s}\, \widetilde{\mathbf{H}}_e \mathbf{x}_{di} + \mathbf{w}_{si}}_{\mathbf{w}'_{si}}, \tag{3.85}$$

where $\mathbf{H}_e = \widetilde{\mathbf{H}}_s - \widehat{\mathbf{H}}_s$ is the estimation error for $\widetilde{\mathbf{H}}_s$, and $\mathbf{w}'_{si}$ is the equivalent noise that has zero mean and the autocorrelation

$$\begin{aligned} \mathbf{R}_{w_s} &= c_\delta N_s\, \mathbb{E}\big[\mathbf{H}_e \mathbf{H}_e^H\big] + \mathbf{I}_{N_s} && (3.86) \\ &= \big(1 + \frac{c_\delta N_d N_s}{1 + c_\tau N_s}\big) \mathbf{I}_{N_s} && (3.87) \end{aligned}$$

Therefore, the mutual information

$$\begin{aligned} I(\mathbf{y}_{si}; \mathbf{x}_{di} | \widehat{\mathbf{H}}_s) &\geq \log\det\left(\mathbf{I}_{N_s} + \frac{c_\delta N_s\, \widehat{\mathbf{H}}_s \widehat{\mathbf{H}}_s^H}{1 + c_\delta N_d N_s / (1 + c_\tau N_s)}\right) && (3.88) \\ &= \log\det\big(\mathbf{I}_{N_d} + \rho_d \overline{\mathbf{H}}_s \overline{\mathbf{H}}_s^H\big), && (3.89) \end{aligned}$$

where $\overline{\mathbf{H}}_s = \frac{1}{\delta_{\hat{H}_s}} \widehat{\mathbf{H}}_s$ is the normalized channel estimate. For the static user, the effective SNR for decoding the dynamic signal is identical to that of the dynamic user.

We now assume the signal for the dynamic user is encoded via a sufficiently long period so that the static user also experiences many channel realizations over this period. Write $\overline{\mathbf{H}}_s = [\overline{\mathbf{H}}_{s1}; \overline{\mathbf{H}}_{s2}]$, where $\overline{\mathbf{H}}_{s1} \in \mathcal{C}^{N_d \times N_d}$ and $\overline{\mathbf{H}}_{s2} \in \mathcal{C}^{(N_s - N_d) \times N_d}$. Then,

$$\begin{aligned} \mathbb{E}\big[I(\mathbf{y}_{si}; \mathbf{x}_{di} | \widehat{\mathbf{H}}_s)\big] &\geq \mathbb{E}\left[\log\det\left(\mathbf{I}_{N_d} + \rho_d\big(\overline{\mathbf{H}}_{s1} \overline{\mathbf{H}}_{s1}^H + \overline{\mathbf{H}}_{s2} \overline{\mathbf{H}}_{s2}^H\big)\right)\right] && (3.90) \\ &\geq \mathbb{E}\left[\log\det\left(\mathbf{I}_{N_d} + \rho_d \overline{\mathbf{H}}_{s1} \overline{\mathbf{H}}_{s1}^H\right)\right], && (3.91) \end{aligned}$$

where in the last equality we use the fact that $\log\det(\mathbf{A} + \mathbf{B}) \geq \log\det \mathbf{A}$ for any positive definite matrices $\mathbf{A}, \mathbf{B}$. Note that $\overline{\mathbf{H}}_{s1}$ has the same distribution as $\overline{\mathbf{H}}_d$, the static user is able to decode the data of the dynamic user.

Using SVD, $\mathbf{X}_d = \mathbf{U}_d \mathbf{\Sigma}_d \mathbf{V}_d^H$, where $\mathbf{U}_d \in \mathcal{C}^{N_d \times N_d}$, $\mathbf{V}_d \in \mathcal{C}^{T_d \times N_d}$ are unitary matrices, and $\mathbf{\Sigma}_d = \text{diag}(\lambda_1, \cdots, \lambda_{N_d})$. Then, we have

$$\begin{aligned} \mathbf{Y}'_s &= \mathbf{Y}_s \mathbf{V}_d \mathbf{\Sigma}_d^{-1} && (3.92) \\ &= \mathbf{H}_s \mathbf{X}_s \mathbf{U}_d + \mathbf{W}_s \mathbf{V}_d \mathbf{\Sigma}_d^{-1} && (3.93) \\ &\triangleq \mathbf{H}_s \mathbf{X}'_s + \mathbf{W}'_s \mathbf{\Sigma}_d^{-1}, && (3.94) \end{aligned}$$

where $\mathbf{X}'_s = \mathbf{X}_s \mathbf{U}_d, \mathbf{W}'_s = \mathbf{W}_s \mathbf{V}_d$. Because $\mathbf{U}_d$, $\mathbf{V}_d$ are unitary, the entries of $\mathbf{X}'_s, \mathbf{W}'_s \in \mathcal{C}^{N_s \times N_d}$ remain i.i.d. $\mathcal{CN}(0,1)$. Define $\mathbf{y}'_s = \mathbf{vec}(\mathbf{Y}'_s)$, $\mathbf{x}'_s = \mathbf{vec}(\mathbf{X}'_s)$, $\mathbf{H}'_s = \mathbf{I}_{N_d} \otimes \mathbf{H}_s$ and

$$\mathbf{w}'_s = \mathbf{vec}(\mathbf{W}'_s \mathbf{\Sigma}_d^{-1}) = \begin{bmatrix} \frac{1}{\lambda_1} \mathbf{w}'_{s1} \\ \vdots \\ \frac{1}{\lambda_{N_d}} \mathbf{w}'_{sN_d} \end{bmatrix}, \tag{3.95}$$

where $\mathbf{w}'_{si}$ is the column i of $\mathbf{W}'_s$. Then, from (3.94), we write $\mathbf{y}'_s \in \mathcal{C}^{N_d N_s \times 1}$ as

$$\mathbf{y}'_s = \mathbf{H}'_s \mathbf{x}'_s + \mathbf{w}'_s. \tag{3.96}$$

The mutual information

$$\begin{aligned} I(\mathbf{Y}_s; \mathbf{X}_s | \mathbf{H}_s, \mathbf{X}_d) &= I(\mathbf{y}'_s; \mathbf{x}'_s | \mathbf{H}_s, \mathbf{X}_d) && (3.97) \\ &= \log\det \left(\mathbf{I}_{N_d N_s} + \mathbf{R}_{w'_s}^{-1} \ \mathbf{H}'_s \mathbf{H}'^H_s \right), && (3.98) \end{aligned}$$

where $\mathbf{R}_{w'_s} = \mathbb{E}[\mathbf{w}'_s \mathbf{w}'^H_s]$ is the noise autocorrelation matrix that is given by

$$\mathbf{R}_{w'_s} = \begin{bmatrix} \mathbb{E}[\lambda_1^{-2}]\mathbf{I}_{N_s} & & \\ & \ddots & \\ & & \mathbb{E}[\lambda_{N_d}^{-2}]\mathbf{I}_{N_s} \end{bmatrix}. \tag{3.99}$$

Therefore, the average rate attained by the static user is

$$\begin{aligned} R_s &= \frac{1}{T_d} \mathbb{E}[I(\mathbf{Y}_s; \mathbf{X}_s | \mathbf{H}_s, \mathbf{X}_d)] && (3.100) \\ &= \frac{1}{T_d} \mathbb{E}\left[\sum_{i=1}^{N_d} \log\det \left(\mathbf{I}_{N_s} + \frac{1}{\mathbb{E}[\lambda_i^{-2}]} \mathbf{H}_s \mathbf{H}_s^H \right) \right] && (3.101) \\ &= \frac{N_d}{T_d} \mathbb{E}\left[\log\det \left(\mathbf{I}_{N_s} + \frac{1}{\mathbb{E}[\lambda_1^{-2}]} \mathbf{H}_s \mathbf{H}_s^H \right) \right], && (3.102) \end{aligned}$$

where the last equality holds because the marginal distributions of λ_i are identical.

CHAPTER 4
CAPACITY LIMITS OF MULTIUSER MULTIANTENNA SPECTRUM SHARING NETWORKS

4.1 Introduction

This chapter studies performance limits of an underlay cognitive network consisting of multiuser and multi-antenna primary and secondary systems. The primary and secondary systems are subject to mutual interference, where the secondary must comply with a set of interference constraints imposed by the primary. We are interested in the secondary throughput, i.e., the sum rate averaged over channel realizations, as the number of secondary users grows. Moreover, we study how the secondary throughput is affected by the size of primary network as well as the severity of the interference constraints.

A summary of the results of this chapter is as follows. We assume that the primary and secondary have N and n users, respectively, and their base stations have M and m antennas, respectively. In this chapter, n is allowed to grow (to infinity) while N, M and m are bounded (not scaling with n).

- **Secondary uplink (MAC):** The secondary throughput is shown to grow as $\Theta(\log n)$, which is achieved by a threshold-based user selection rule. More precisely, the throughput of the secondary MAC channel grows as $\frac{m}{N_p+1}\log n + O(1)$ when it coexists with the primary broadcast channel, and grows as $\frac{m}{M_p+1}\log n + O(1)$ when it coexists with the primary MAC channel. By developing asymptotically tight upper bounds, these growth rates are further proven to be optimal. Moreover, the interference on the primary system can be asymptotically forced to zero, while the secondary throughput still grows as $\Theta(\log n)$. Specifically, for some non-negative exponent q, the interference on the primary can be made to decline as $\Theta(n^{-q})$, while the throughput of a secondary MAC grows as $\frac{m-qN_p}{N_p+1}\log n+O(1)$ and $\frac{m-qM_p}{M_p+1}\log n+O(1)$, respectively in cases of primary broadcast and MAC channel. The above results imply that asymptotically the secondary system can attain a non-trivial throughput without degrading the performance of the primary system.

- **Secondary downlink (broadcast):** The secondary throughput is shown to scale with $m\log\log n+O(1)$ in the presence of either the primary broadcast or MAC channel. Hence, the growth rate of throughput is unaffected (thus optimal) by the presence of the primary system. In addition, the interference on the primary can be asymptotically forced to zero, while maintaining the secondary throughput as $\Theta(\log\log n)$. Specifically, for an arbitrary exponent $0<q<1$, the interference can be made to decline as $\Theta\big((\log n)^{-q}\big)$, while the secondary throughput grows as $m(1-q)\log\log n+O(1)$.

- **Non-homogeneous networks:** Secondary throughput under non-homogeneous inter-node link gains is studied for both secondary MAC and broadcast. It is shown that even if the nodes experience unequal path loss and shadowing, under a broad class of path loss and shadowing models, the secondary throughput growth rates remain unaffected.

Much of the past work in the underlay cognitive radio involves point-to-point primary and secondary systems. Ghasemi and Sousa [35] studies the ergodic capacity of a point-to-point secondary link under various fading channels. Multiple antennas at the secondary transmitter are exploited by [36] to manage the tradeoff between the secondary throughput and the interference on the primary. In the context of multi-user cognitive radios, Zhang et al. [37] studies the power allocation of a single-antenna secondary system under various transmit power constraints as well as interference constraints. Gastpar [38] studies the secondary capacity via translating a receive power constraint into a transmit power constraint.

Recently, ideas from opportunistic communication [39] were used in underlay cognitive radios by selectively activating one or more secondary users to maximize the secondary throughput while satisfying interference constraints. The user selection in cognitive radio is complicated because the secondary system must be mindful of two criteria: The interference on the primary and the rate provided to the secondary. Karama et al. [40] selects secondary users with channels almost orthogonal to a single primary user, so that the interference on the primary is reduced. Jamal et al. [41, 42] obtains interesting scaling results for the throughput by selecting users causing the least interference. Some distinctions of our work and [41,42] are worth noting. First, Jamal et al. [41,42] studies the hardening of sum rate via convergence in probability, while we analyze the throughput, which requires a very different approach.[1] Second, we study a multi-antenna cognitive network and the effect of the primary network size (number of constraints) on the secondary throughput, whereas [41, 42] consider a single antenna network with a single primary constraint.

We use the following notation: $[\,\cdot\,]_{i,j}$ refers to the (i, j) element in a matrix, $|\cdot|$ refers to the cardinality of a set or the Euclidean norm of a vector, $\mathrm{diag}(\cdot)$ refers to a diagonal matrix, $\mathrm{tr}(\cdot)$ refers to the trace of a matrix, and I_k refers to the $k \times k$ identity matrix. All $\log(\cdot)$ is natural base. For any $\epsilon > 0$, some positive c_1 and c_2, and sufficiently large n:

$$\begin{aligned} f(n) = O\big(g(n)\big) &: & |f(n)| &< c_1\,|g(n)|, \\ f(n) = \Theta\big(g(n)\big) &: & c_2\,|g(n)| < |f(n)| &< c_1\,|g(n)|, \\ f(n) = o\big(g(n)\big) &: & |f(n)| &< \epsilon\,|g(n)|. \end{aligned}$$

In this chapter, we define throughput as the sum rate averaged over all channel realizations. We let $\mathcal{R}^{opt}_{mac,w/o}$ and $\mathcal{R}^{opt}_{bc,w/o}$ be the *maximum* throughput achieved by the secondary MAC and broadcast channel *in the absence of* the primary, respectively. In this case, we

[1] In general, convergence in probability does not imply convergence in any moment (thus average throughput) [43]. For example, consider a sequence of rates $R_n = \log(1+X_n)$, where

$$X_n = \begin{cases} 1 & \text{with probability } 1 - \frac{1}{n} \\ \exp(n^2) & \text{with probability } \frac{1}{n} \end{cases}.$$

Then, $\lim_{n\uparrow\infty} R_n = \log 2$ in probability, however, $\lim_{n\uparrow\infty} \mathbb{E}[R_n] = \infty$ in probability. Therefore, the average rate $\mathbb{E}[R_n]$ may not be predicted based on the hardening (in probability) of R_n.

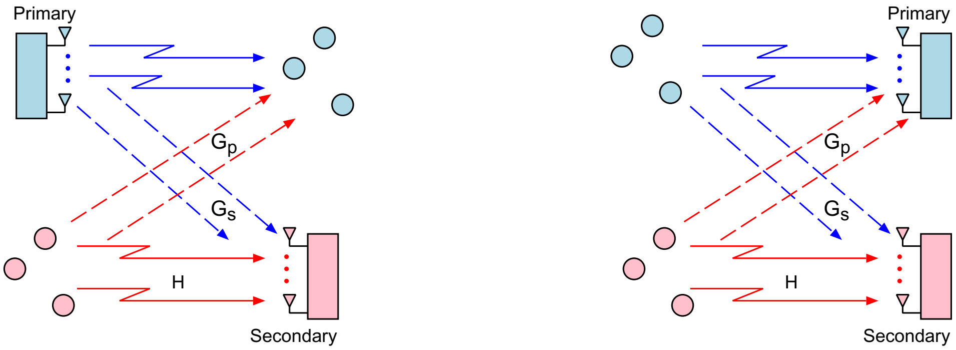

Figure 4.1. Coexistence of the secondary MAC channel and the primary system

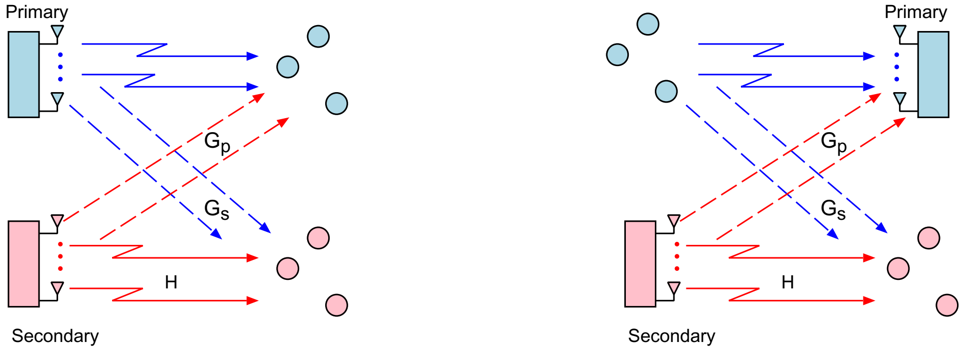

Figure 4.2. Coexistence of the secondary broadcast channel and the primary system

have regular MAC and broadcast channels, and it is well known that $\mathcal{R}^{opt}_{mac,w/o}$ scales as $m \log n$ [26], and $\mathcal{R}^{opt}_{bc,w/o}$ scales as $m \log \log n$ [34].

The remainder of this chapter is organized as follows. Section 4.2 describes the system model. The throughput of the secondary MAC channel is studied in Section 4.3, where in Section 4.3.3 we prove the achieved throughout is asymptotically optimal. The average throughput of the secondary broadcast channel is investigated in Section 4.4. Section 4.5 studies the effect of path-loss and shadowing on the secondary throughput. Numerical results are shown in Section 4.6.

4.2 System Model

We consider a cognitive network consisting of a primary and a secondary, each being either a MAC or broadcast channel (Figure 4.1 and Figure 4.2). The primary system has one base station with M_p antennas and N_p users, while the secondary system consists of one base station with m antennas and n users. The primary and secondary are subject to mutual interference, which is treated as noise (no interference decoding). The secondary system must comply with a set of interference power constraints imposed by the primary. For simplicity of exposition, primary and secondary users are assumed initially to have one antenna, however, as shown in the sequel, most of the results can be directly extended to a scenario where each user has multiple antennas.

A block-fading channel model is assumed. All channel coefficients are fixed throughout each transmission block, and are independent, identically distributed (i.i.d.) circularly-symmetric-complex-Gaussian with zero mean and unit variance, denoted by $\mathcal{CN}(0,1)$. The secondary base station acts as a scheduler: For each transmission block, a subset of the secondary users is selected to transmit to (or receive from) the secondary base station. We denote the collection of selected (active) secondary users as $\mathcal{S}$.

We begin by introducing a system model that applies to all four scenarios in Figures 4.1 and 4.2, thus simplifying notation in the remainder of the chapter. The secondary received signal is given by:

$$\mathbf{y} = \mathbf{H}(\mathcal{S})\,\mathbf{x}_s + \mathbf{G}_s\,\mathbf{x}_p + \mathbf{w}, \tag{4.1}$$

where $\mathbf{y}$ represents the received signal vector, either signals at a multi-antenna base station (uplink) or at different users (downlink). $\mathbf{H}(\mathcal{S})$ is the channel coefficient matrix between the active secondary users and their base station. $\mathbf{G}_s$ represents the cross channel coefficient matrix from the primary transmitter(s) to the secondary receiver(s). The primary and secondary transmit signal vectors are $\mathbf{x}_p$ and $\mathbf{x}_x$. The variable $\mathbf{w}$ is the received noise vector, where each entry of $\mathbf{w}$ is i.i.d. $\mathcal{CN}(0,1)$.

We assume both primary and secondary systems use Gaussian signaling, subject to short-term power constraints. The transmit covariance matrices of the primary and secondary systems are

$$Q_p = \mathbb{E}\big[\mathbf{x}_p \mathbf{x}_p^\dagger\big], \tag{4.2}$$

and

$$Q_s = \mathbb{E}\big[\mathbf{x}_s \mathbf{x}_s^\dagger\big]. \tag{4.3}$$

When the secondary is a MAC channel, each secondary user is subject to an individual short term power constraint ρ_s. The users do not cooperate, therefore Q_s is diagonal:

$$Q_s = \mathrm{diag}\big(\rho_1, \cdots, \rho_{|\mathcal{S}|}\big), \tag{4.4}$$

where $\rho_\ell \le \rho_s$, for $\ell = 1, \cdots, |\mathcal{S}|$. In this case, $\mathbf{H}(\mathcal{S})$ has dimension $m \times |\mathcal{S}|$.

When the secondary is a broadcast channel, we assume the secondary base station is subject to a short term power constraint P_s:

$$\mathrm{tr}(Q_s) \le P_s. \tag{4.5}$$

In this case, $\mathbf{H}(\mathcal{S})$ has dimension $|\mathcal{S}| \times m$.

When the primary is a MAC channel, each primary user transmits with power ρ_p without user cooperation:

$$Q_p = \rho_p \, I_{N_p}. \tag{4.6}$$

Furthermore, each receive antenna at the primary base station can tolerate interference with power Γ from the secondary system,[2] that is

$$\big[\mathbf{G}_p \, Q_s \, \mathbf{G}_p^\dagger\big]_{\ell,\ell} \le \Gamma, \tag{4.7}$$

for $\ell = 1, \cdots, M_p$, where $\mathbf{G}_p$ represents the cross channel coefficient matrix from the secondary base station (or active users) to the primary base station.

[2] If each primary antenna or user tolerates a different interference power, the results of this chapter still hold, as seen later.

When the primary is a broadcast channel, the power constraint at the primary base station is $\mathrm{tr}(Q_p) \leq P_p$. For simplicity, we assume[3]

$$Q_p = \frac{P_p}{M_p} I_{M_p}. \tag{4.8}$$

Furthermore, each primary user tolerates interference with power Γ:

$$\left[\mathbf{G}_p \, Q_s \, \mathbf{G}_p^\dagger\right]_{\ell,\ell} \leq \Gamma, \tag{4.9}$$

for $\ell = 1, \cdots, N_p$, where $\mathbf{G}_p$ is the cross channel coefficient matrix from the secondary base station (or active users) to the primary users.

4.3 Cognitive MAC Channel

Consider a MAC secondary in the presence of either a broadcast or MAC primary. We wish to find how much throughput is available to the secondary subject to rigid constraints on the secondary-on-primary interference. We first construct a transmission strategy and find the corresponding (achievable) throughput. Then, we develop upper bounds that are tight with respect to the throughput achieved.

The framework for the transmission strategy is as follows: For each transmission block, the secondary base station determines an active user set $\mathcal{S}$ as well as transmit power for all active users Q_s. For each transmission, from (4.1), the sum rate of the secondary system is [44]:

$$R_{mac} = \log\det\left(I + \mathbf{H}(\mathcal{S})Q_s\mathbf{H}^\dagger(\mathcal{S}) + \mathbf{G}_s Q_p \mathbf{G}_s^\dagger\right) - \log\det\left(I + \mathbf{G}_s Q_p \mathbf{G}_s^\dagger\right). \tag{4.10}$$

subject to the interference constraints (4.9) and (4.7) for the primary broadcast and MAC channel respectively.

The secondary throughput is obtained by averaging R_{mac} over channel realizations

$$\mathcal{R}_{mac} = \mathbb{E}[R_{mac}]. \tag{4.11}$$

[3] The asymptotic results remain the same, even if we allow Q_p to be an arbitrary covariance matrix.

For the development of upper bounds, we assume the secondary base station knows all the channels. This is a genie-like argument that is used solely for development of upper bounds. For the achievable scheme, the requirement is more modest and is outlined after the description of the achievable scheme (see Remark 4.3.1).

4.3.1 Achievable Scheme

The objective is to choose $\mathcal{S}$ and Q_s, i.e., the secondary active transmitters and their power, such that secondary throughput is maximized subject to interference constraints on the primary.

The choice of $\mathcal{S}$ and Q_s is coupled through the interference constraints: Either more secondary users can transmit with smaller power, or fewer of them with higher power. We focus on a simple power policy that all active secondary users transmit with the maximum allowed power ρ_s. Hence, given an active user set $\mathcal{S}$, we have

$$Q_s = \rho_s I_{|\mathcal{S}|}. \tag{4.12}$$

It will be shown that the on-off transmission (without any further power adaptation) suffices to asymptotically achieve the maximum throughput. Furthermore, its simplicity facilitates analysis.

Recall that each primary user can tolerate interference with power Γ. The interference on a primary user is guaranteed to be below this level if k_s secondary users are active, each causing interference no more than $\alpha = \frac{\Gamma}{k_s}$. This bound allows us to honor the interference constraints on the primary while decoupling the action of different secondary users. Based on this observation, we construct a user selection rule as follows. First, we define an eligible secondary user set that disqualifies users that cause too much interference on the primary:

$$\mathcal{A} = \big\{ i : \rho_s \big| [\mathbf{G}_p]_{ji} \big|^2 < \alpha, \text{ for all } j \big\}, \tag{4.13}$$

where $[\mathbf{G}_p]_{ji}$ is the channel coefficient from the secondary user i to the primary user (antenna) j, and α is a pre-designed interference quota. A secondary user is eligible if its interference on each primary user (antenna) is less than α. Now, to satisfy the interference bound, we limit the number of secondary transmitters to no more than k_s, where

$$k_s = \frac{\Gamma}{\alpha}. \tag{4.14}$$

If $|\mathcal{A}| \leq k_s$, then all eligible users can transmit. If $|\mathcal{A}| > k_s$, then k_s users will be chosen *randomly* from among the eligible users to transmit.[4] The number of eligible users, $|\mathcal{A}|$, is a random variable; the number of active users is

$$|\mathcal{S}| = \min\big(k_s, |\mathcal{A}|\big). \tag{4.15}$$

The transmission of $|\mathcal{S}|$ eligible users induces interference no more than Γ on any primary user or antenna. Notice that the manner of user selection guarantees that the channel coefficients in $\mathbf{H}(\mathcal{S})$ remain independent and distributed as $\mathcal{CN}(0,1)$.

Now we want to design an interference quota α to maximize the secondary throughput. Neither very small nor very large values of α are useful within our framework: If α is very small, for most transmissions few (if any) secondary users will be eligible, thus the secondary throughput will be small. If $\alpha > \Gamma$, any transmitting user might violate the interference constraint, so the secondary must shut down (equivalently, we have $k_s < 1$). The value of individual interference constraint α, or equivalently k_s, must be set somewhere between these extremes.

Clearly, a desirable outcome would be to allow exactly the number of users that are indeed eligible for transmission, i.e., $k_s \approx |\mathcal{A}|$. But one cannot guarantee this in advance because $|\mathcal{A}|$ is a random variable. Motivated by this general insight, we choose α such that

$$k_s = \mathbb{E}[|\mathcal{A}|]. \tag{4.16}$$

In Section 4.3.3, we will verify that this choice of α is enough to asymptotically achieve the maximum throughput.

[4] Naturally the number of active users must be an integer, i.e., $\lfloor k_s \rfloor$. We do not carry the floor operation in the following developments for simplicity, noting that due to the asymptotic nature of the analysis, the floor operation has no effect on the final results.

Remark 4.3.1 *The above scheme does not require the secondary users to have full channel knowledge. Each secondary user can compare its own cross channel gains with a pre-defined interference quota α, and then decide its eligibility. After this, each eligible user can inform the secondary base station via 1-bit, so that the secondary base station can determine $\mathcal{A}$ without knowing the cross channels from the secondary users to the primary system. The secondary channels $\mathbf{H}(\mathcal{S})$ and the cross channels $\mathbf{G}_s$ can be estimated at the secondary base station. Therefore, this scheme can be implemented with little exchange of channel knowledge.*

4.3.2 Throughput Calculation

Secondary MAC with Primary Broadcast

The primary base station transmits to N_p primary users, where each user tolerates interference with power Γ. Notice that in (4.13), $[\mathbf{G}_p]_{ji}$ is the channel coefficient from the secondary user i to the primary user j which is i.i.d. $\mathcal{CN}(0,1)$. Thus, $\left|[\mathbf{G}_p]_{ji}\right|^2$ is i.i.d. exponential. Therefore, $|\mathcal{A}|$ is binomially distributed with parameter (n,p), where

$$p = \left(1 - e^{-\frac{\alpha}{\rho_s}}\right)^{N_p}. \tag{4.17}$$

From (4.16), the interference quota $\alpha = \frac{\Gamma}{k_s}$ is chosen such that

$$\begin{aligned} k_s &= np \\ &= n\big(\frac{\Gamma}{\rho_s}\big)^N k_s^{-N} + O\big(nk_s^{-(N+1)}\big). \end{aligned} \tag{4.18}$$

Denote the associated solution for k_s as $\bar{k}_s$:

$$\bar{k}_s = \left(\frac{\Gamma}{\rho_s}\right)^{\frac{N_p}{N_p+1}} (n)^{\frac{1}{N_p+1}} + O(1). \tag{4.19}$$

Thus, we can see $\Theta(n^{\frac{1}{N_p+1}})$ secondary users are allowed to transmit, and the interference quota is on the order of $\Theta(n^{-\frac{1}{N_p+1}})$. With the above choice of interference quota, or the number of allowable active users, we state one of the main results of this chapter as follows.

Theorem 4.3.1 *Consider a secondary MAC with a m-antenna base station and n users each with power constraint ρ_s. The secondary MAC operates in the presence of a primary broadcast channel with an M-antenna transmitter with power P_p to N users each with interference tolerance Γ. The secondary throughput satisfies:*

$$\mathcal{R}_{mac} \geq \frac{m}{N_p+1}\log n + \frac{1}{N_p+1}\log\left(\rho_s \Gamma^{N_p}\right) - m\log(1+P_p) + O\big(n^{-\frac{1}{N_p+1}}\log n\big), \tag{4.20}$$

$$\mathcal{R}_{mac} \leq \frac{m}{N_p+1}\log n + \frac{1}{N_p+1}\log\left(\rho_s \Gamma^{N_p}\right) - \mathcal{R}_I + O\big(n^{-\frac{1}{N_p+1}}\big).$$

with

$$\mathcal{R}_I = m_{\mathsf{min}}\log\left(1 + \frac{P_p}{M_p}\exp\left(\frac{1}{m_{\mathsf{min}}}\sum_{j=1}^{m_{\mathsf{min}}}\sum_{i=1}^{m_{\mathsf{max}}-j}\frac{1}{i} - \gamma\right)\right), \tag{4.21}$$

where $m_{\mathsf{min}} = \min(m, M_p)$, $m_{\mathsf{max}} = \max(m, M_p)$ and γ is the Euler's constant. This throughput is achieved under the threshold-based user selection with the choice of $\bar{k}_s$ given by (4.19).

Proof See Section 4.7.1.

Remark 4.3.2 *The essence of the above result is that the secondary throughput grows as $\frac{m}{N_p+1}\log n + O(1)$, which implies that the secondary throughput decreases almost linearly with the number of primary constraints as $n \to \infty$. A noteworthy special case is when the primary base station chooses to transmit to a number of users equal to the number of its transmit antennas ($N_p = M_p$), a strategy which is known to be near-optimum in terms of sum rate [13]. Under this condition:*

$$\mathcal{R}_{mac} = \frac{m}{M_p+1}\log n + O(1).$$

Therefore, we have

$$\lim_{n\to\infty}\frac{\mathcal{R}_{mac}}{\mathcal{R}^{opt}_{mac,w/o}} = \frac{1}{M_p+1}, \tag{4.22}$$

where $\mathcal{R}^{opt}_{mac,w/o}$ is the maximum throughput of the secondary MAC in the absence of *the primary system. This ratio shows that the* compliance penalty *of the secondary MAC system and its relationship with the characteristics of the primary network.*

It is noteworthy that although Γ does not affect the growth rate, it is an important parameter. Both the lower and upper bounds have the term $\frac{1}{N+1}\log(\rho_s\Gamma^N) = \frac{1}{N+1}\log\rho_s + \frac{N}{N+1}\log\Gamma$, thus throughput is an increasing function of Γ. One can also see that the interference tolerance Γ is more important than secondary power ρ_s, respectively by a factor of $\frac{N}{N+1}$ versus $\frac{1}{N+1}$.

Remark 4.3.3 *The results in Theorem 4.3.1 can be directly extended to a scenario where each primary user tolerates a different level of interference. As long as all primary users allow non-zero interference (no matter how small), we can let Γ be the minimum allowable interference, and the theorem still holds.*

So far we have analyzed the effect of small but constant primary interference constraints, and shown that the secondary throughput improves with increasing the number of secondary users. The flexibility provided by the increasing number of secondary users can be exploited not only to increase secondary throughput, but also to reduce the primary interference. In fact, it is possible to simultaneously suppress the interference on the primary down to *zero* while increasing the secondary throughput proportional to $\log n$. The following corollary makes this idea precise:

Corollary 4.3.2 *Assuming the interference on each primary user is bounded as $\Theta(n^{-q})$, the secondary throughput satisfies*

$$\mathcal{R}_{mac} = \frac{m - qN_p}{N_p + 1}\log n + O(1), \tag{4.23}$$

where $0 < q < \frac{1}{N_p}$.

Proof Because the proof of Theorem 4.3.1 holds for $\Gamma = \Theta(n^{-q})$, the corollary follows by substituting $\Gamma = \Theta(n^{-q})$ into the lower and upper bounds given by Theorem 4.3.1.

Remark 4.3.4 *The corollary above explores a tradeoff where primary interference is made to decrease polynomially, i.e., proportional to n^{-q}. We saw that this leads to a secondary* *throughput that decreases linearly in q. If we reduce the primary interference more slowly, e.g., decreasing as $\Theta(1/\log n)$, one can verify that $\mathcal{R}_{mac} = \frac{m}{N+1}\log n - \frac{N}{N+1}\log\log n + O(1)$, which achieves the optimal growth rate even though the throughput is reduced. Conversely, if we try to suppress the primary interference faster than $\Theta(n^{-1/N})$, the secondary throughput will asymptotically remain stagnant, i.e., $\mathcal{R}_{mac} = o(\log n)$, since in this case $\bar{k}_s = O(1)$ according to* (4.19).

Secondary MAC with Primary MAC

Recall that each antenna at the primary base station allows interference with power Γ. By regarding each antenna of the primary base station as a virtual user, we can re-use most of the analysis that was developed in the previous section. Thus, the steps leading to Eq. (4.19) can be repeated to obtain the number of allowable active secondary users:

$$\bar{k}_s = \left(\frac{\Gamma}{\rho_s}\right)^{\frac{M_p}{M_p+1}} (n)^{\frac{1}{M_p+1}} + O(1). \tag{4.24}$$

With this allowable active users $\bar{k}_s$ and slight modifications, we obtain a result that parallels Theorem 4.3.1.

Theorem 4.3.3 *Consider a secondary MAC with an m-antenna base station and n users each with power constraint ρ_s. The secondary MAC operates in the presence of a primary MAC channel where each user transmits with power ρ_p to an M_p-antenna base station with interference tolerance Γ on each antenna. The secondary throughput satisfies:*

$$\mathcal{R}_{mac} \geq \frac{m}{M_p+1}\log n + \frac{1}{M_p+1}\log\left(\rho_s\Gamma^{M_p}\right) - m\log(1+\rho_p N_p) + O\big(n^{-\frac{1}{M_p+1}}\log n\big), \tag{4.25}$$

$$\mathcal{R}_{mac} \leq \frac{m}{M_p+1}\log n + \frac{1}{M_p+1}\log\left(\rho_s\Gamma^{M_p}\right) - \mathcal{R}_I + O\big(n^{-\frac{1}{M_p+1}}\big), \tag{4.26}$$

with

$$\mathcal{R}_I = m_{\mathsf{min}}\log\left(1+\rho_p\exp\left(\frac{1}{m_{\mathsf{min}}}\sum_{j=1}^{m_{\mathsf{min}}}\sum_{i=1}^{m_{\mathsf{max}}-j}\frac{1}{i}-\gamma\right)\right), \tag{4.27}$$

where $m_{\mathsf{min}} = \min(m, N_p)$, $m_{\mathsf{max}} = \max(m, N_p)$ and γ is the Euler's constant. This throughput is achieved under the threshold-based user selection with the choice of $\bar{k}_s$ given by (4.24).

A tradeoff exists between the primary interference reduction and the secondary throughput enhancement, which is stated by the following corollary. Remark 4.3.4 is again applicable here.

Corollary 4.3.4 *Assuming the interference on each antenna of the primary base station is bounded as $\Theta(n^{-q})$, the secondary throughput satisfies*

$$\mathcal{R}_{mac} = \frac{m - qM_p}{M_p + 1} \log n + O(1), \tag{4.28}$$

where $0 < q < \frac{1}{M_p}$.

4.3.3 Upper Bounds for Secondary Throughput

So far we have seen achievable rates of a cognitive MAC channel in the presence of either a primary broadcast or MAC. We now develop corresponding upper bounds.

Theorem 4.3.5 *Consider a secondary MAC with an m-antenna base station and n users. The* maximum *throughput of the secondary, $\mathcal{R}_{mac}^{opt}$, satisfies*

$$\mathcal{R}_{mac}^{opt} \leq \frac{m}{N_p + 1} \log n + O(\log \log n), \tag{4.29}$$

in the presence of a primary broadcast channel transmitting to N_p users. Similarly, $\mathcal{R}_{mac}^{opt}$ satisfies

$$\mathcal{R}_{mac}^{opt} \leq \frac{m}{M_p + 1} \log n + O(\log \log n), \tag{4.30}$$

in the presence of a primary MAC, where each user transmits to an M_p-antenna base station.

Proof See Section 4.7.2.

Remark 4.3.5 *By comparing the upper bounds with the achievable rates obtained by the thresholding strategy, we see that the achievable rates are at most $O(\log \log n)$ away from the upper bounds, a difference which is negligible relative to the dominant term $\Theta(\log n)$. Thus, the growth of the* maximum *throughput of a cognitive MAC is $\frac{m}{N_p+1} \log n$ in the presence of* *the primary broadcast channel, and* $\frac{m}{M_p+1}\log n$ *in the presence of the primary MAC channel. Both the achievable rates and the upper bounds show that the average cognitive sum-rate decreases almost linearly with the number of primary-imposed constraints, asymptotically.*

4.3.4 Discussion

Recall that our method determines eligible cognitive MAC users based on their cross channel gains. To satisfy the interference constraints, our selection rule then allows $\Theta(n^{\frac{1}{N_p+1}})$, or $\Theta(n^{\frac{1}{M_p+1}})$, of these users to be active simultaneously, in the presence of either the primary broadcast or MAC. If there are more eligible users than the allowed number, we choose from among the eligible users randomly. In this process, the forward channel gain of the cognitive users does not come into play, and still an optimal growth rate is achieved. This can be intuitively explained as follows. The total received signal power at the cognitive base station grows linearly with the number of active users, and the total received signal power determines the throughput. On the other hand, selecting good cognitive users according to their secondary channel strengths can only offer logarithmic power gains (with respect to n) [39], which is negligible compared to the linear gain due to increasing the number of active users. Therefore the cross channel gains are more important in this case. Note that we do not imply that knowledge of the cognitive forward channel is useless; our conclusion only says that once the cross channels are taken into account, the *asymptotic growth* of the secondary throughput cannot be improved by any use of the cognitive forward channel.

Although we have allowed the base stations to have multiple antennas, so far the users have been assumed to have only one antenna. We now consider a generalization to the case where all users have multiple antennas. Consider a secondary MAC in the presence of a primary broadcast, where each primary and secondary user have t_p and t_s antennas respectively. We apply a separate interference constraint on each antenna of each primary user, which guarantees the satisfaction of the overall interference constraint on any primary user. On each of the t_s-antenna secondary users, we shall allocate $t_s - 1$ degrees of freedom for zero-forcing and only one degree of freedom for cognitive transmission. Using this strategy, we can ensure that $t_s - 1$ of the primary receive antennas are exempt from interference. Thus, the total number of interference constraints will reduce from t_pN_p to $t_pN_p + 1 - t_s$. By using an analysis similar to the development of Theorem 4.3.1, one can show that the growth rate $\frac{m \log n}{\max(1, t_pN_p+2-t_s)}$ is achievable. For the converse, the situation is more complicated, because here the correlation among the antennas of the secondary users must be accounted for. Nevertheless, in some cases it is possible to show without much difficulty that the above achieved throughput is indeed asymptotically optimal. For example, in the presence of the primary MAC, if $t_s > M_p$, the secondary MAC channel can have a throughput that grows as $m \log n$ by letting each active secondary user completely eliminate the interference on the primary. Similarly, in the presence of a primary broadcast channel, if $t_s > t_pN_p$, the secondary MAC channel can also have a throughput that grows as $m \log n$. The achieved growth rate is optimal because it coincides with the the growth rate of $\mathcal{R}^{opt}_{mac,w/o}$, which is always an upper bound.

4.4 Cognitive Broadcast Channel

4.4.1 Achievable Scheme

We consider a random beam-forming technique where the secondary base station opportunistically transmits to m secondary users simultaneously [34]. Specifically, the secondary base station constructs m orthonormal beams, denoted by $\{\phi_j\}_{j=1}^m$, and assigns each beam to a secondary user. Then, the secondary base station broadcasts to m selected users. The selection of users and beam assignment will be addressed shortly.

Considering an equal power allocation among m users, the transmitted signal from the secondary base station is given by:

$$\mathbf{x}_s = \sum_{j=1}^{m} \sqrt{\frac{P}{m}}\, \phi_j \, x_j, \tag{4.31}$$

where ϕ_j is the beam-forming vector j with dimension $m \times 1$, x_j is the signal transmitted along with the beam j, and P is the total transmit power. In this case, we have

$$Q_s = \frac{P}{m} I_m. \tag{4.32}$$

Notice that P is subject to the power constraint P_s as well as a set of interference constraints imposed by the primary. Thus, the value of P depends on the cross channels from the secondary base station to the primary system.

Assuming the beam j is assigned to the user i. From (4.1) and (4.31), the received signal at the secondary user i is given by

$$y_i = \mathbf{h}_i^\dagger \phi_j x_j + \sum_{k \neq j} \mathbf{h}_i^\dagger \phi_k x_k + \mathbf{g}_{s,i}^\dagger \mathbf{x}_p + w_i, \tag{4.33}$$

where $\mathbf{h}_i^\dagger$ is the $1 \times m$ vector of channel coefficients from the secondary base station to the secondary user i, and $\mathbf{g}_{s,i}^\dagger$ is the $1 \times M_p$ (or $1 \times N_p$) vector of channel coefficients from the primary base station (or users) to the secondary user i. The received signal-to-noise-plus-interference-ratio (SINR) at the secondary user i (with respect to the beam j) is

$$\mathsf{SINR}_{i,j} = \frac{\frac{P}{m}|\mathbf{h}_i^\dagger \phi_j|^2}{1 + \frac{P}{m}\sum_{k \neq j} |\mathbf{h}_i^\dagger \phi_k|^2 + \mathbf{g}_{s,i}^\dagger \, Q_p \, \mathbf{g}_{s,i}}. \tag{4.34}$$

The random beam technique assigns each beam to the secondary user that results in the highest SINR. Because the probability of more than two beams being assigned to the same secondary user is negligible [34], we have the secondary throughput

$$\mathcal{R}_{bc} \approx \mathbb{E}\left[\sum_{j=1}^{m} \log\left(1 + \max_{1 \leq i \leq n} \mathsf{SINR}_{i,j}\right)\right] \tag{4.35}$$

$$= m\mathbb{E}\left[\log\left(1 + \max_{1 \leq i \leq n} \mathsf{SINR}_{i,j}\right)\right]. \tag{4.36}$$

The above analysis holds in the presence of either the primary broadcast or MAC channel; the only difference is the constraints on P and Q_p. Since the SINR is symmetric across all beams, the subscript j will be omitted in the following analysis.

Remark 4.4.1 *We briefly address the issue of channel state information. All users are assumed to have receiver side channel state information. On the transmit side, the secondary base station only needs to know SINR and does not need to have full channel knowledge. Each secondary user can estimate its own SINR with respect to each beam, and feed it back to the secondary base station [34]. Based on collected SINR, the secondary base station performs user selection. The secondary base station needs to know* $\mathbf{G}_p$ *to adjust* P *such that the interference constraints on the primary are satisfied.*

4.4.2 Throughput Calculation

Secondary Broadcast with Primary Broadcast

The secondary system has to comply with the constraints on N_p primary users. To maximize the throughput, the secondary base station transmits at the maximum allowable power. From (4.9) and (4.32), we have

$$P = \min\big(\frac{m\Gamma}{|\mathbf{g}_{p,1}^\dagger|^2}, \cdots, \frac{m\Gamma}{|\mathbf{g}_{p,N_p}^\dagger|^2}, P_s\big), \tag{4.37}$$

where $\mathbf{g}_{p,\ell}^\dagger$ is the row ℓ of $\mathbf{G}_p$. Then, we substitute Q_p given by (4.8) into (4.34), and obtain the SINR at the secondary user i with respect to the beam j:

$$\mathsf{SINR}_i = \frac{|\mathbf{h}_i^\dagger \phi_j|^2}{\frac{m}{P} + \sum_{k\neq j} |\mathbf{h}_i^\dagger \phi_k|^2 + \frac{mP_p}{M_pP}|\mathbf{g}_{s,i}|^2}. \tag{4.38}$$

Our analysis of $\max_i \mathsf{SINR}_i$, which is required to evaluate the throughput in Eq. (4.36), does not follow [34] because the denominator involves a sum of two Gamma distributions with different scale parameters: $\sum_{k\neq j} |\mathbf{h}_i^\dagger \phi_k|^2$ has Gamma$(m-1,1)$ and $\frac{mP_p}{M_pP}|\mathbf{g}_{s,i}|^2$ has Gamma$(M_p, \frac{mP_p}{M_pP})$. Fortunately, lower and upper bounds can be leveraged to simplify the analysis. We define:

$$\theta = \frac{mP_p}{M_pP}. \tag{4.39}$$

We consider the case when $\frac{mP_p}{M_pP_s} \geq 1$. The techniques can be generalized to the case of $\frac{mP_p}{M_pP_s} < 1$.[5] When $\frac{mP_p}{M_pP_s} \geq 1$, we have $\theta \geq 1$ for all P. We define:

$$L_i = \frac{|\mathbf{h}_i^\dagger \phi_j|^2}{\frac{m}{P} + \theta\big(\sum_{k \neq j} |\mathbf{h}_i^\dagger \phi_k|^2 + |\mathbf{g}_{s,i}|^2\big)}, \tag{4.40}$$

and

$$U_i = \frac{|\mathbf{h}_i^\dagger \phi_j|^2}{\frac{m}{P} + \theta |\mathbf{g}_{s,i}|^2}, \tag{4.41}$$

where L_i and U_i are random variables that depend on channel realizations. Conditioned on P, the denominators of L_i and U_i have Gamma distributions, which simplifies the analysis.

For $1 \leq i \leq n$, we have

$$L_i \leq \mathsf{SINR}_i \leq U_i. \tag{4.42}$$

Hence, for any channel realization,

$$L_{max} \leq \max_{1 \leq i \leq n} \mathsf{SINR}_i \leq U_{max}, \tag{4.43}$$

where $L_{max} = \max_i L_i$ and $U_{max} = \max_i U_i$. Therefore, the secondary throughput is bounded as follows:

$$m\mathbb{E}\big[\log(1 + L_{max})\big] \leq \mathcal{R}_{bc} \leq m\mathbb{E}\big[\log(1 + U_{max})\big]. \tag{4.44}$$

We study the lower and upper bounds given by (4.44), instead of directly analyzing $\mathcal{R}_{bc}$. Some useful properties of L_{max} and U_{max} are as follows.

Lemma 4.4.1 *Conditioned on $P = \rho$,*

$$P_r\left(L_{max} \geq b_n - \frac{\rho}{m}\log\log n \,\middle|\, P = \rho\right) = 1 - \Theta\left(\frac{1}{n}\right), \tag{4.45}$$

$$P_r\left(U_{max} < d_n + \frac{\rho}{m}\log\log n \,\middle|\, P = \rho\right) = 1 - \Theta\left(\frac{1}{\log n}\right), \tag{4.46}$$

$$\mathbb{E}\left[U_{max} \,\middle|\, U_{max} > d_n + \frac{\rho}{m}\log\log n, P = \rho\right] < O(n \log n), \tag{4.47}$$

[5] When $\frac{mP_p}{M_pP_s} < 1$, one can define $\theta = \max(\frac{mP_p}{M_pP}, 1)$. Then, we can use Bayesian expansion via conditioning on $\{P < \frac{mP_p}{M_p}\}$ and its complement, where both conditional terms can be shown to have the same growth rate.

where $b_n = \frac{\rho}{m}\log n - \frac{\rho(m+M_p-1)}{m}\log\log n + O\big(\log\log\log n\big)$ *and* $d_n = \frac{\rho}{m}\log n - \frac{\rho M_p}{m}\log\log n + O\big(\log\log\log n\big)$*.*

Proof See Section 4.7.3.

Based on the above two lemmas, we obtain the following results for the secondary throughput.

Theorem 4.4.2 *Consider a secondary broadcast channel with* n *users and an* m*-antenna base station with power constraint* P_s*. The secondary broadcast operates in the presence of a primary broadcast channel transmitting with power* P_p *to* N_p *users each with interference tolerance* Γ*. The secondary throughput satisfies:*

$$\mathcal{R}_{bc} > m\log\big(\Gamma\log n\big) - m\log\big(\tilde{\mu}_1 + \frac{m\Gamma}{P_s}\big) + O\big(\frac{\log\log n}{\log n}\big),$$
$$\mathcal{R}_{bc} < m\log(\Gamma\log n) - m\log\tilde{\mu}_2 + O(1),$$

where $\tilde{\mu}_1 = \mathbb{E}[\max_{1\le i\le N_p}|\mathbf{g}_{p,i}^{\dagger}|^2]$ *and* $\tilde{\mu}_2 = \big(\mathbb{E}\big[1/\max_{1\le i\le N_p}|\mathbf{g}_{p,i}^{\dagger}|^2\big]\big)^{-1}$*.*

Proof See Section 4.7.4.

Remark 4.4.2 *The result above states that* $\mathcal{R}_{bc} = m\log\log n + O(1)$*, thus*

$$\lim_{n\to\infty}\frac{\mathcal{R}_{bc}}{\mathcal{R}_{bc,w/o}^{opt}} = 1, \tag{4.48}$$

where $\mathcal{R}_{bc,w/o}^{opt}$ *is the maximum throughput of the secondary broadcast channel in the absence of the primary system. Therefore, the achieved throughput is asymptotically optimal, because we always have* $\mathcal{R}_{bc} \le \mathcal{R}_{bc,w/o}^{opt}$*. Thus, we have a positive result: The growth rate of the secondary throughput is unaffected by the constraints and interference imposed by the primary, as long as each primary user tolerates some fixed interference* Γ*.*

The above results naturally lead to the question: How small can we make the interference on the primary, while still having a secondary throughput that grows as $\Theta(\log\log n)$. We find that Γ, the interference on each primary user, can asymptotically go to zero, as shown by the next corollary.

Corollary 4.4.3 *Assuming the interference on each primary user is bounded as $\Theta\big((\log n)^{-q}\big)$, the secondary throughput satisfies:*

$$\mathcal{R}_{bc} = (1-q)m\log\log n + O(1), \tag{4.49}$$

where $0 < q < 1$.

Remark 4.4.3 *This result sheds lights on the tradeoff between two goals of a cognitive radio system: High throughput for the secondary and low interference on the primary. For primary interference reduction up to $\Theta\big((\log\log n)^{-1}\big)$, one can verify that $\mathcal{R}_{bc} = m\log\log n - m\log\log\log n + O(1)$, which still achieves the double logarithmic growth rate for secondary throughput. It is possible to reduce the interference faster than $\Theta((\log n)^{-1})$, but this will make $\mathcal{R}_{bc} = o(\log\log n)$.*

Remark 4.4.4 *It can be shown that the growth rate of the secondary throughput does not depend on the transmit covariance Q_p of the primary . To see this, we decompose $Q_p = U\Lambda U^\dagger$, where U is an unitary matrix and $\Lambda = diag(\lambda_1, \cdots, \lambda_M)$, $0 < \lambda_1 \leq \cdots \leq \lambda_M < P_p$. From (4.34), we have $\mathbf{g}_{s,i}^\dagger\, Q_p\, \mathbf{g}_{s,i} = \tilde{\mathbf{g}}_{s,i}^\dagger\, \Lambda\, \tilde{\mathbf{g}}_{s,i}$, where $\tilde{\mathbf{g}}_{s,i} = U^\dagger \mathbf{g}_{s,i}$ has the same distribution as $\mathbf{g}_{s,i}$ [21]. Therefore, $\lambda_1|\tilde{\mathbf{g}}_{s,i}|^2 \leq \mathbf{g}_{s,i}^\dagger\, Q_p\, \mathbf{g}_{s,i} \leq \lambda_M|\tilde{\mathbf{g}}_{s,i}|^2$. With the exception of a slightly different definition of θ, the analysis for $Q_p = I$ will follow.*

Secondary Broadcast with Primary MAC

The analysis of this case closely parallels the analysis of the primary broadcast. The secondary transmit power is given by

$$P = \min\big(\frac{m\Gamma}{|\mathbf{g}_{p,1}^\dagger|^2}, \cdots, \frac{m\Gamma}{|\mathbf{g}_{p,M_p}^\dagger|^2}, P_s\big), \tag{4.50}$$

where $\mathbf{g}_{p,\ell}^\dagger$ is the row ℓ of $\mathbf{G}_p$. The MAC primary system produces power $N_p\rho_p$ and has M_p interference constraints. From the viewpoint of the secondary, this is all the information that is needed. Therefore the analysis of Theorem 4.4.2 can be essentially repeated to obtain the following result.

Theorem 4.4.4 *Consider a secondary broadcast channel with n users and an m-antenna base station with power constraint P_s. The secondary broadcast operates in the presence of a primary MAC where each user transmits with power ρ_p to an M_p-antenna base station with interference tolerance Γ on each antenna. The secondary throughput satisfies:*

$$\mathcal{R}_{bc} > m\log\big(\Gamma\log n\big) - m\log\big(\tilde{\mu}_3 + \frac{m\Gamma}{P_s}\big) + O\big(\frac{\log\log n}{\log n}\big)$$
$$\mathcal{R}_{bc} < m\log(\Gamma\log n) - m\log\tilde{\mu}_4 + O(1),$$

where $\tilde{\mu}_3 = \mathbb{E}[\max_{1\le i\le M_p}|\mathbf{g}_{p,i}^{\dagger}|^2]$ and $\tilde{\mu}_4 = \big(\mathbb{E}\big[1/\max_{1\le i\le M_p}|\mathbf{g}_{p,i}^{\dagger}|^2\big]\big)^{-1}$.

Remark 4.4.5 *Theorem 4.4.2 and Theorem 4.4.4 can be extended to a scenario where each primary (secondary) user has multiple antennas via regarding each primary and secondary antenna as a virtual user. Using analysis similar to the single-antenna case, the secondary broadcast channel can be shown to achieve a throughput scaling as $m\log\log n$ (thus optimal). The details are straightforward and are therefore omitted for brevity.*

Similar to Corollary 4.4.3, we can also obtain the tradeoff between the primary interference reduction and the secondary throughput enhancement as follows. All the remarks following Corollary 4.4.3 apply to the present case as well.

Corollary 4.4.5 *Assuming the interference on each antenna of the primary base station is bounded as $\Theta\big((\log n)^{-q}\big)$, the secondary throughput satisfies:*

$$\mathcal{R}_{bc} = (1-q)m\log\log n + O(1) \tag{4.51}$$

where $0 < q < 1$.

4.5 Capacity Scaling under Path Loss and Shadowing

The results so far were developed assuming all fading channels obey the same distribution, i.e., for a homogeneous network. In this section, we generalize our results by allowing different users to experience varying path loss and shadowing. We consider the combined effect of path loss and shadowing as a multiplicative factor on the channel gain. The probabilistic behavior of this multiplicative factor can in general be complicated because it depends on the spatial distribution of users, whose randomness will induce a distribution on path loss, as well as the composition of the terrain. However, certain assumptions can be made about it from first principles. We assume the support of the probability density of path loss and shadowing is positive[6] and bounded. This is equivalent to saying that the distance between nodes cannot be arbitrarily large or arbitrarily small, and that shadowing attenuates but is not a perfect isolator of emissions [45]. Conditioned on a realization of path loss and shadowing, the resulting fading coefficient is assumed to be a Rayleigh random variable whose variance is determined by the value of path loss and shadowing.

In this section we concentrate on a broadcast primary. Similar results hold with little variation for the primary MAC channel, and are omitted for brevity.

Our basic idea of characterizing the secondary throughput in the presence of path loss and shadowing is as follows. We find an upper (lower) bound on the secondary throughput by constructing a homogeneous network whose throughput is no larger (smaller) than the throughput under any realization of path loss and shadowing. The throughput of the homogeneous networks that bound our performance follows the analysis of previous sections. We then show the scaling of the throughput lower and upper bounds are identical.

4.5.1 Secondary MAC

A *homogeneous* secondary MAC channel with cross link variance ν and secondary link variance μ can be shown, using methods of the previous sections, to have a throughput charac-

[6] For ease of exposition we have assumed that under path loss and shadowing a link is not completely lost. It is possible to carry through the analysis as long as at least $O(n)$ secondary links remain available, and no more than $o(\log n)$ cross links go to zero. If too many cross links disappear due to path loss and shadowing, effectively that part of the network is no longer cognitive and the nature of the problem would be changed.

terized by:

$$\begin{aligned}\mathcal{R}^{iid}_{mac}(\nu,\mu,n) \geq \frac{m}{N_p+1}\log n + \frac{1}{N_p+1}\log\big(\rho_s(\frac{\Gamma}{\nu})^{N_p}\big) + \log\mu \\ - m\log(1+P_p) + O\big(n^{-\frac{1}{N_p+1}}\log n\big), \\ \mathcal{R}^{iid}_{mac}(\nu,\mu,n) \leq \frac{m}{N_p+1}\log n + \frac{1}{N_p+1}\log\big(\rho_s(\frac{\Gamma}{\nu})^{N_p}\big) + \log\mu \\ - \mathcal{R}_I + O\big(n^{-\frac{1}{N_p+1}}\log n\big). \end{aligned} \tag{4.52}$$

Now, consider a *heterogeneous* network where path loss and shadowing of the nodes vary according to a distribution with positive and bounded support. Then, conditioned on the path loss and shadowing, the channel coefficient from the secondary user i to the primary user j is $\mathcal{CN}(0,\sigma^2_{ji})$, and from the secondary user i to all the co-located secondary base station antennas is $\mathcal{CN}(0,\sigma^2_{si})$, while all other channel coefficients are i.i.d. $\mathcal{CN}(0,1)$. Let $X = \{\sigma^2_{ji},\, \sigma^2_{si},\, 1 \leq j \leq N,\, 1 \leq i \leq n\}$, the set of all random channel variances. The positivity and boundedness assumptions for the support of the path loss distribution are formalized by $0 < \nu_1 \leq \sigma^2_{ji} \leq \nu_2$ and $0 < \mu_1 \leq \sigma^2_{si} \leq \mu_2$.

We now outline an argument based on the intuition that the secondary throughput increases or at worst stays the same if one secondary link improves, and that the secondary throughput does not increase if one cross link gets stronger. To make this argument precise, the cross-link variance σ^2_{ji} can be absorbed into the interference constraint Γ, resulting in an equivalent cross link with unit variance and interference constraint $\frac{\Gamma}{\sigma^2_{ji}}$. So a stronger cross link is equivalent to a stricter interference constraint, therefore the secondary throughput is non-increasing in σ^2_{ji}. Similarly, the secondary link variance σ^2_j can be absorbed into the secondary transmit power for the secondary throughput calculation, leading to an effective transmit power $\sigma^2_j\rho_s$ over a link of unit variance. Thus, the throughput is non-decreasing with σ^2_j.

Based on the above argument, we always have $\mathcal{R}_{mac} \geq \mathcal{R}^{iid}_{mac}(\nu_2,\mu_1,n)$ and $\mathcal{R}_{mac} \leq \mathcal{R}^{iid}_{mac}(\nu_1,\mu_2,n)$, because $\nu_1 \leq \sigma^2_{ji} \leq \nu_2$ and $\mu_1 \leq \sigma^2_{si} \leq \mu_1$ for any realization of X. Therefore,

$$\mathbb{E}[\mathcal{R}_{mac}] \geq \mathcal{R}^{iid}_{mac}(\nu_2,\mu_1,n), \tag{4.53}$$

and

$$\mathbb{E}[\mathcal{R}_{mac}] \leq \mathcal{R}_{mac}^{iid}(\nu_1, \mu_2, n). \tag{4.54}$$

From (4.52), we conclude that the growth rate of our proposed technique under path loss and shadowing is $\frac{m}{N+1}\log n + O(1)$.

However, this has not fully settled the capacity question, because the upper bound was calculated only under a specific scheme. A stronger upper bound is obtained by noting that for any transmission scheme, the throughput of the heterogenous network is smaller than the throughput of the homogeneous network with cross link variance ν_1 and secondary link variance μ_2. The latter throughput can be shown, following the analysis of Theorem 4.3.5, to be upper bounded by $\frac{m}{N+1}\log n + O(\log\log n)$.

Thus, we have lower and upper bounds whose order matches, and we have the following result.

Theorem 4.5.1 *Consider a secondary MAC channel with n users, m antennas at the base station, and power constraint ρ_s, in the presence of a primary that broadcasts with power constraint P_p to N users with interference constraint Γ. The users are randomly located resulting in path loss and shadowing coefficients whose combined effect can be characterized by a random variable whose support is over a strictly positive bounded interval, then:*

$$\frac{m}{N_p+1}\log n + O(1) \leq \; \mathbb{E}\big[\mathcal{R}_{mac}^{opt}\big] \; \leq \frac{m}{N_p+1}\log n + O(\log\log n). \tag{4.55}$$

Therefore, the throughput grows with $\frac{m}{N_p+1}\log n$.

4.5.2 Secondary Broadcast

Now, we consider a secondary broadcast channel. A *homogeneous* secondary broadcast channel with primary-to-secondary channel variance ν and secondary link variance μ can be shown to have a throughput

$$\mathcal{R}_{bc}^{iid}(\nu, \mu, n) = m\log\big(\mu\Gamma\log n\big) + O(1). \tag{4.56}$$

Consider the channel coefficient from all the co-located secondary base-station antennas to the secondary user i to be $\mathcal{CN}(0, \sigma_i^2)$, and the primary to the secondary user i to be $\mathcal{CN}(0, \sigma_{pi}^2)$, while all other channel coefficients are i.i.d. $\mathcal{CN}(0, 1)$. Let $X = \{\sigma_i^2, \sigma_{pi}^2, 1 \leq i \leq n\}$, where $0 < \mu_3 \leq \sigma_i^2 \leq \mu_4$ and $0 < \nu_3 \leq \sigma_{pi}^2 \leq \nu_4$.

Similar to the argument for cognitive MAC channel, $\mathcal{R}_{bc}$ decreases with σ_{pi}^2 but increases with σ_i^2. Therefore, we have

$$\mathcal{R}_{bc}^{iid}(\nu_4, \mu_3, n) \leq \mathbb{E}_X[\mathcal{R}_{bc}] \leq \mathcal{R}_{bc}^{iid}(\nu_3, \mu_4, n). \tag{4.57}$$

Thus, in the presence of path loss and shadowing, we have the following result.

Theorem 4.5.2 *Consider a secondary broadcast channel with n users, m antennas at the base station, and power constraint ρ_s, in the presence of a primary that broadcasts with power constraint P_p to N users with interference constraint Γ. The users are randomly located resulting in path loss and shadowing coefficients whose combined effect can be characterized with a random variable whose support is over a strictly positive bounded interval, then:*

$$\lim_{n\to\infty} \frac{\mathbb{E}\left[\mathcal{R}_{bc}^{opt}\right]}{m \log\log n} = 1 \tag{4.58}$$

Remark 4.5.1 *The heterogeneity of the following channels does not affect the throughput growth rate in a straightforward manner, thus it is not considered in the above analysis: (1) For the secondary MAC channel, the cross channel between the primary and secondary base-stations, whose variance only affects the interference on the secondary (independent of n), and (2) for the secondary broadcast channel, the cross channel from the secondary base-station to the primary, which only affects the secondary transmit power that is once again independent of n.*

4.6 Numerical Results

In this section, we concentrate on numerical results in the presence of the primary broadcast channel and the results in the presence of the primary MAC channel are similar thus omitted.

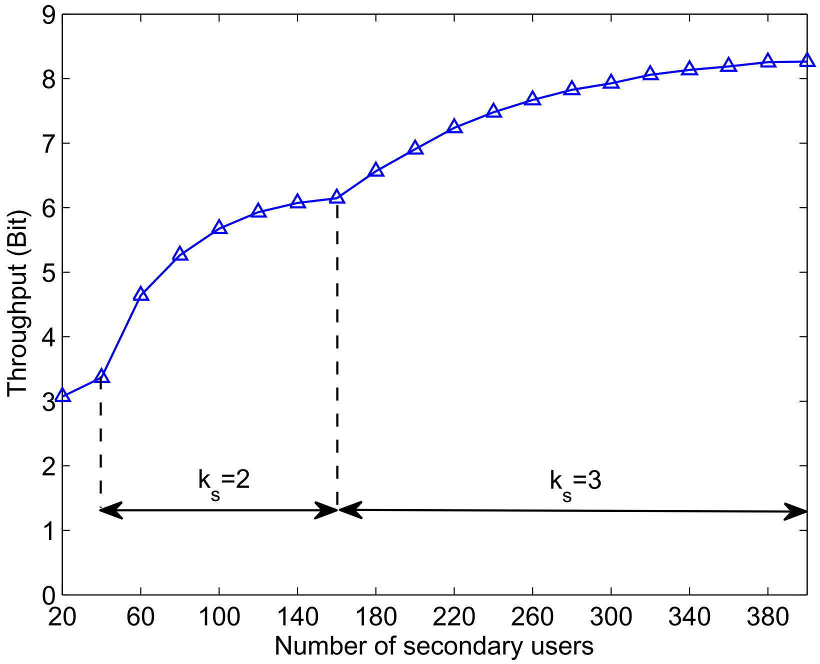

Figure 4.3. Secondary MAC: Throughput versus user number ($\Gamma = 2$)

For all simulations, we consider: $P_p = P_s = \rho_s = 5$, the secondary base station has $m = 4$ antennas, and the primary base station has $M_p = 2$ antennas and the number of primary users is $N_p = 2$.

Figure 4.3 illustrates the secondary throughput given by Theorem 4.3.1. The allowable interference power on each primary user is $\Gamma = 2$. The slope of the throughput curve is discontinuous at some points, because the allowable number of active secondary users must be an integer $\lfloor k_s \rfloor$ (also see Eq.(4.18)). As mentioned earlier, the floor operation does not affect the asymptotic results. Figure 4.4 presents the tradeoff between the tightness of the primary constraints and the secondary throughput, as shown by Corollary 4.3.2. The interference power constraint Γ is $2n^{-q}$ for $q = 0.1$ and 0.2 respectively. As expected, for $q = 0.2$ the interference on primary decreases faster than $q = 0.1$ and the secondary throughput increases more slowly.

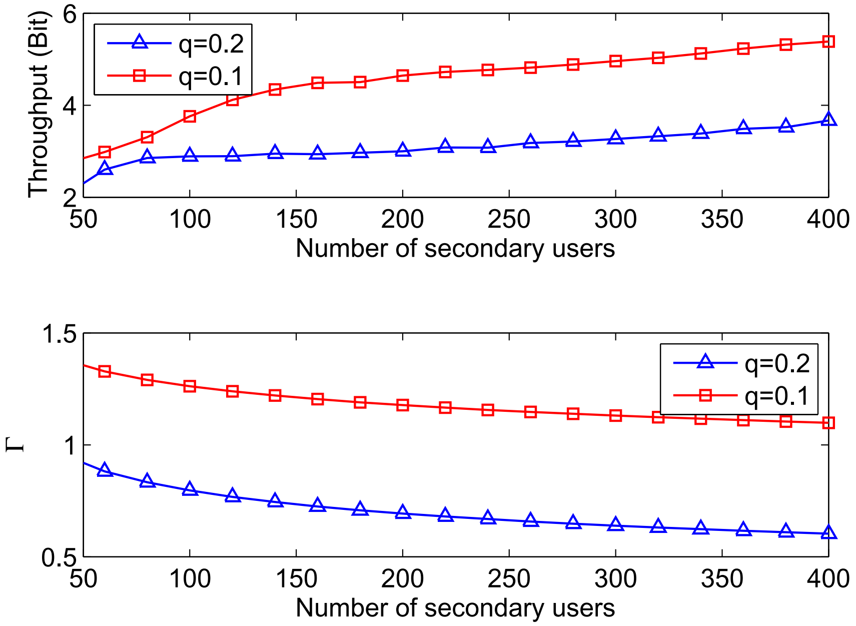

Figure 4.4. Secondary MAC: Throughput versus user number ($\Gamma = 2n^{-q}$)

Figure 4.5 shows the secondary throughput versus the number of secondary users in the presence of the primary broadcast channel (Theorem 4.4.2), where the interference power is $\Gamma = 2$. In Figure 4.6, we show the tradeoff between the secondary throughput and the interference on the primary, as described in Corollary 4.4.3. We set Γ to decline as $2(\log n)^{-q}$, for $q = 0.5$ and $q = 0.8$, respectively. Clearly, for $q = 0.5$, the interference power decreases faster than $q = 0.8$, while the secondary throughput increases more slowly.

4.7 Proof of Theorem and Lemma

4.7.1 Proof of Theorem 4.3.1

We rewrite (4.10) as

$$R_{mac} = \log\det\left(I + \mathbf{H}(\mathcal{S})Q_s\mathbf{H}^{\dagger}(\mathcal{S})\big(I + \mathbf{G}_s Q_p \mathbf{G}_s^{\dagger}\big)^{-1}\right). \tag{4.59}$$

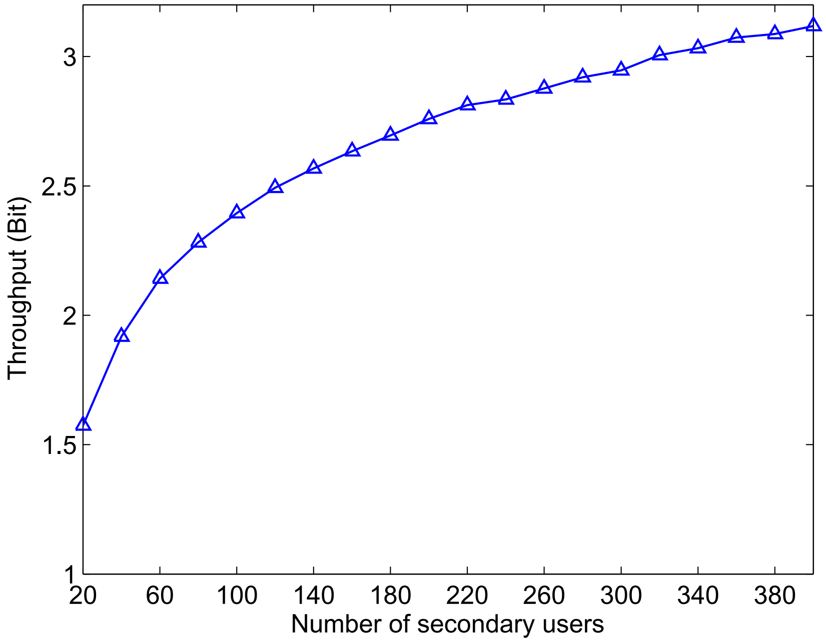

Figure 4.5. Secondary broadcast: Throughput versus user number ($\Gamma = 2$)

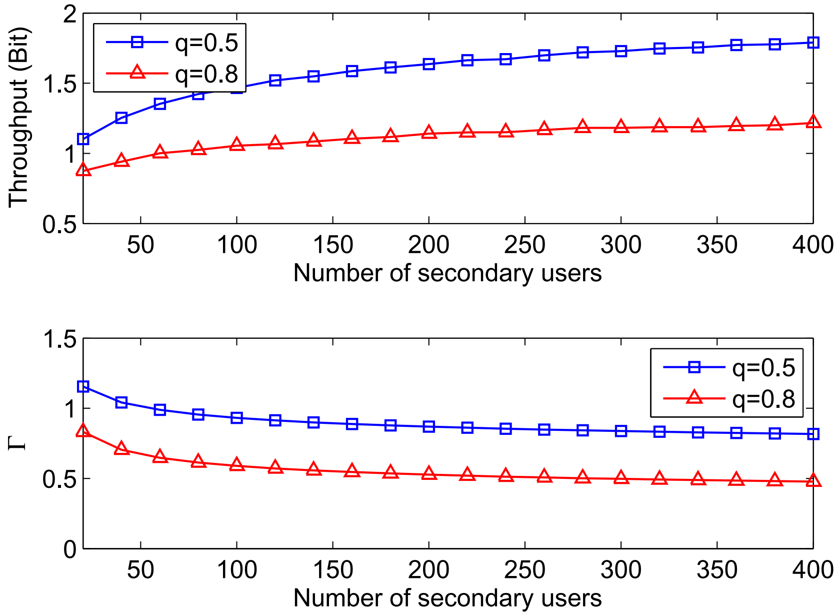

Figure 4.6. Secondary broadcast: Throughput versus user number ($\Gamma = 2(\log n)^{-q}$)

The secondary throughput is calculated by averaging the instant rate R_{mac} over all channel realizations, i.e., $\mathbf{H}$ and $\mathbf{G}_s$, thus

$$\begin{aligned}\mathcal{R}_{mac} &= \mathbb{E}_{\mathbf{H}}\big[\mathbb{E}_{\mathbf{G}_s}[R_{mac}\,|\,\mathbf{H}]\big] \\ &= \mathbb{E}_{\mathbf{H}}\bigg[\mathbb{E}_{\mathbf{G}_s}\bigg[\log\det\Big(I+\mathbf{H}(\mathcal{S})Q_s\mathbf{H}^\dagger(\mathcal{S})\times\big(I+\mathbf{G}_sQ_p\mathbf{G}_s^\dagger\big)^{-1}\Big)\bigg]\bigg]. \qquad (4.60)\end{aligned}$$

Because for any positive definite matrix A and B, the function $\log\det(I+AB^{-1})$ is convex in B [46, Lemma II.3], we apply the Jensen inequality on the right hand side of the inequality (4.60), i.e., taking expectation with respect to $\mathbf{G}_s$

$$\begin{aligned}\mathcal{R}_{mac} &> \mathbb{E}_{\mathbf{H}}\bigg[\log\det\Big(I+\mathbf{H}(\mathcal{S})Q_s\mathbf{H}^\dagger(\mathcal{S})\times\big(I+\mathbb{E}[\mathbf{G}_sQ_p\mathbf{G}_s^\dagger]\big)^{-1}\Big)\bigg] \qquad (4.61)\\ &= \mathbb{E}_{\mathbf{H}}\bigg[\log\det\Big(I+\frac{\rho_s}{1+P_p}\mathbf{H}(\mathcal{S})\mathbf{H}^\dagger(\mathcal{S})\Big)\bigg], \qquad (4.62)\end{aligned}$$

where in (4.62) we use the facts that $Q_p = P_pI/M$ and $\mathbb{E}[\mathbf{G}_s\mathbf{G}_s^\dagger] = M_pI_m$ since each entry of $\mathbf{G}_s$ is i.i.d. $\mathcal{CN}(0,1)$.

Now we bound the right hand side of (4.62). Recall that $|\mathcal{A}|$ and $|\mathcal{S}|$ are the random number of eligible users and active users, respectively. By the Chebychev inequality, for any $\epsilon > 0$, we have

$$\begin{aligned}P_r\Big(|\mathcal{A}| > (1-\epsilon)\bar{k}_s\Big) &> 1-\frac{1-p}{\epsilon^2 np} \qquad (4.63)\\ &= 1-O\big(\bar{k}_s^{-1}\big), \qquad (4.64)\end{aligned}$$

where in the above we use the fact $\bar{k}_s = np$. Then, we expand (4.62) based the event $\{|\mathcal{A}| > (1-\epsilon)\bar{k}_s\}$ and its complement, and discard the non-negative term associated with its complement:

$$\begin{aligned}&\mathcal{R}_{mac}\\ &> \mathbb{E}\bigg[\log\det\Big(I+\frac{\rho_s}{1+P_p}\mathbf{H}(\mathcal{S})\mathbf{H}^\dagger(\mathcal{S})\Big)\,\bigg|\,|\mathcal{A}| > (1-\epsilon)\bar{k}_s\bigg]\times P_r\Big(|\mathcal{A}| > (1-\epsilon)\bar{k}_s\Big) \qquad (4.65)\\ &\geq \mathbb{E}\bigg[\log\det\Big(I+\frac{\rho_s}{1+P_p}\mathbf{H}(\mathcal{S})\mathbf{H}^\dagger(\mathcal{S})\Big)\,\bigg|\,|\mathcal{A}| = (1-\epsilon)\bar{k}_s\bigg]\times\Big(1-O\big(\bar{k}_s^{-1}\big)\Big) \qquad (4.66)\\ &= \mathbb{E}\bigg[\log\det\Big(I+\frac{\rho_s}{1+P_p}\mathbf{H}(\mathcal{S})\mathbf{H}^\dagger(\mathcal{S})\Big)\,\bigg|\,|\mathcal{S}| = (1-\epsilon)\bar{k}_s\bigg]\times\Big(1-O\big(\bar{k}_s^{-1}\big)\Big), \qquad (4.67)\end{aligned}$$

where in the inequality (4.66), we apply the result in (4.64) and the fact that the conditional expectation of the right hand side of (4.65) is non-decreasing in $|\mathcal{A}|$. Since $|\mathcal{S}| = (1-\epsilon)\bar{k}_s$ in case of $|\mathcal{A}| = (1-\epsilon)\bar{k}_s$, then we obtain (4.67) due to the throughput depending on $|\mathcal{A}|$ via the size of $\mathcal{S}$.

Recall that each entry of $\mathbf{H}(\mathcal{S})$ is i.i.d. $\mathcal{CN}(0,1)$. Conditioned on $|\mathcal{S}| = (1-\epsilon)\bar{k}_s$, $\mathbf{H}(\mathcal{S})\mathbf{H}^\dagger(\mathcal{S})$ is a Wishart Matrix with degrees of freedom $(1-\epsilon)\bar{k}_s$, we have [47, Lemma A]

$$\begin{aligned}
\mathcal{R}_{mac} &> \Big(m\log\big(1+\frac{(1-\epsilon)\rho_s\bar{k}_s}{1+P_p}\big)+O\big(\bar{k}_s^{-1}\big)\Big)\times\Big(1-O\big(\bar{k}_s^{-1}\big)\Big) &(4.68)\\
&= m\log\big(1+P_p+(1-\epsilon)\rho_s\bar{k}_s\big)+O\big(\frac{\log\bar{k}_s}{\bar{k}_s}\big)-m\log(1+P_p) &(4.69)\\
&= m\log\rho_s\bar{k}_s+m\log(1-\epsilon)-m\log(1+P_p)+O\big(\frac{\log\bar{k}_s}{\bar{k}_s}\big), &(4.70)
\end{aligned}$$

where in (4.70) we use the identity $\log(x+y) = \log x + \log(1+y/x)$, for $x, y > 0$. Since the strict inequality (4.70) holds for any $\epsilon > 0$, thus $\log(1-\epsilon) < 0$ but can be arbitrarily close to zero, by the definition of inequality we have

$$\mathcal{R}_{mac} \geq m\log\rho_s\bar{k}_s - m\log(1+P_p)+O\big(\frac{\log\bar{k}_s}{\bar{k}_s}\big). \tag{4.71}$$

Now we find an upper bound for $\mathcal{R}_{mac}$. For convenience, we denote (see (4.10))

$$R_{mac,0} = \log\det\Big(I+\rho_s\mathbf{H}(\mathcal{S})\mathbf{H}^\dagger(\mathcal{S})+\mathbf{G}_s\,Q_p\,\mathbf{G}_s^\dagger\Big), \tag{4.72}$$

and

$$R_I = \log\det\Big(I+\mathbf{G}_s\,Q_p\,\mathbf{G}_s^\dagger\Big). \tag{4.73}$$

So the throughput can be written as

$$\mathcal{R}_{mac} = \mathbb{E}\big[R_{mac,0}\big]-\mathbb{E}\big[R_I\big]. \tag{4.74}$$

Using the inequality $\det(A) \leq \big(\mathrm{tr}(A)/k)\big)^k$ [17, p. 680], where A is a $k\times k$ positive definite matrix, $R_{mac,0}$ is bounded by

$$R_{mac,0} \leq m\log\Big(1+\frac{1}{m}\mathrm{tr}\Big(\rho_s\mathbf{H}(\mathcal{S})\mathbf{H}^\dagger(\mathcal{S})+\mathbf{G}_s\,Q_p\,\mathbf{G}_s^\dagger\Big)\Big). \tag{4.75}$$

Therefore,

$$\begin{aligned}\mathbb{E}[R_{mac,0}] &\le m\mathbb{E}\Big[\log\Big(1+\frac{1}{m}\mathrm{tr}\Big(\rho_s\mathbf{H}(\mathcal{S})\mathbf{H}^\dagger(\mathcal{S})+\mathbf{G}_s\,Q_p\,\mathbf{G}_s^\dagger\Big)\Big)\Big] &(4.76)\\ &\le m\log\Big(1+\frac{\rho_s}{m}\mathbb{E}\big[\mathrm{tr}\big(\mathbf{H}(\mathcal{S})\mathbf{H}^\dagger(\mathcal{S})\big)\big]+\frac{1}{m}\mathbb{E}\big[\mathrm{tr}\big(\mathbf{G}_sQ_p\mathbf{G}_s^\dagger\big)\big]\Big) &(4.77)\\ &\le m\log\big(1+\rho_s\bar{k}_s+P_p\big), &(4.78)\end{aligned}$$

where (4.77) uses the Jensen inequality. To obtain the inequality (4.78), we use the facts that $\mathbb{E}\big[\mathrm{tr}\big(\mathbf{G}_sQ_p\mathbf{G}_s^\dagger\big)\big]=P_p$ by substituting Q_p given by (4.8) as well as $\mathbb{E}\big[\mathrm{tr}\big(\mathbf{H}(\mathcal{S})\mathbf{H}^\dagger(\mathcal{S})\big)\big]\le m\bar{k}_s$ due to $|\mathcal{S}|\le\bar{k}_s$.

Now we lower bound the second term in (4.74). From [48, Theorem 1], we have

$$\begin{aligned}\mathbb{E}[R_I] &\ge m_{\mathsf{min}}\log\left(1+\frac{P_p}{M_p}\exp\left(\frac{1}{m_{\mathsf{min}}}\sum_{j=1}^{m_{\mathsf{min}}}\sum_{i=1}^{m_{\mathsf{max}}-j}\frac{1}{i}-\gamma\right)\right)\\ &\triangleq \mathcal{R}_I, &(4.79)\end{aligned}$$

where $m_{\mathsf{min}}=\min(m,M_p)$, $m_{\mathsf{max}}=\max(m,M_p)$ and γ is the Euler's constant. Notice that $\mathcal{R}_I$ is a finite constant independent of n and Γ.

Combining (4.78) and (4.79), we have

$$\mathcal{R}_{mac}\le m\log(1+\rho_s\bar{k}_s+P_p)-\mathcal{R}_I. \tag{4.80}$$

Finally, substituting $\bar{k}_s$ given by (4.19) and noting that $\bar{k}_s=\Theta(n^{\frac{1}{N_p+1}})$, we have

$$\begin{aligned}\mathcal{R}_{mac} &\ge \frac{m}{N_p+1}\log n+\frac{1}{N_p+1}\log\big(\rho_s\Gamma^{N_p}\big)-m\log(1+P_p)+O\big(n^{-\frac{1}{N_p+1}}\log n\big) &(4.81)\\ \mathcal{R}_{mac} &\le \frac{m}{N_p+1}\log n+\frac{1}{N_p+1}\log\big(\rho_s\Gamma^{N_p}\big)-\mathcal{R}_I+O\big(n^{-\frac{1}{N_p+1}}\big), &(4.82)\end{aligned}$$

where we use the identity $\log(x+y)=\log x+\log(1+y/x)$, for $x,y>0$, in the above inequalities. This completes the proof.

Remark 4.7.1 *The primary transmit covariance matrix Q_p can be arbitrary and does not affect the growth rate of $\mathcal{R}_{mac}$. For any Q_p, we have $Q_p=U\Lambda U^\dagger$, where U is an unitary and $\Lambda=diag[\lambda_1,\cdots,\lambda_M]$. For the lower bound, in (4.61) we have $\mathbb{E}[\mathbf{G}_sQ_p\mathbf{G}_s^\dagger]=$*

$\mathbb{E}[\mathbf{G}_s U \Lambda U^\dagger \mathbf{G}_s^\dagger] = \mathbb{E}[\mathbf{G}_{s1} \Lambda \mathbf{G}_{s1}^\dagger]$, *where each entry of* $\mathbf{G}_{s1}$ *is still i.i.d.* $\mathcal{CN}(0,1)$ *[21]. Let* $\mathbf{g}_i$ *be the column* i *of* $\mathbf{G}_{s1}$, *then* $\mathbb{E}[\mathbf{G}_s Q_p \mathbf{G}_s^\dagger] = \mathbb{E}[\sum_{i=1}^M \lambda_i \mathbf{g}_i \mathbf{g}_i^\dagger] = \sum_{i=1}^M \lambda_i I_m$. *Since* $tr(Q_p) = P_p$, $\sum_{i=1}^M \lambda_i = P_p$, *which yields the same bound as* (4.62), *and the same development of the lower bound. For the upper bound, we note that in* (4.77) $\mathbb{E}\big[tr\big(\mathbf{G}_s Q_p \mathbf{G}_s^\dagger\big)\big] = tr\big(\mathbb{E}[\mathbf{G}_{s1} \Lambda \mathbf{G}_{s1}^\dagger]\big) = \sum_{i=1}^M \lambda_i tr(I_m) = mP_p$, *which yields the same bound* (4.78) *and thus the development of the upper bound.*

4.7.2 Proof of Theorem 4.3.5

We develop an upper bound for the secondary throughput in the presence of the primary broadcast only; the development is similar in the presence of the primary MAC and thus is omitted. We consider an arbitrary active user set $\mathcal{S}$ and transmit covariance matrix given by (4.4), such that the interference constraints on the primary are satisfied.

By removing the interference from the primary to the secondary, the secondary throughput is enlarged. Then, starting from (4.10) and using the inequality $\det(A) \leq \big(\mathrm{tr}(A)/k\big)^k$ [17, p. 680], where $A_{k\times k}$ is a positive definite matrix, we have

$$R_{mac} \leq m \log\left(1 + \frac{1}{m}\mathrm{tr}\big(\mathbf{H}(\mathcal{S}) Q_s \mathbf{H}^\dagger(\mathcal{S})\big)\right). \tag{4.83}$$

Let $\mathbf{h}_i$ be the $m \times 1$ vector of channel coefficients from the secondary user i ($i \in \mathcal{S}$) to the secondary base station, corresponding to a certain column of $\mathbf{H}(\mathcal{S})$. Since Q_s is diagonal, we have

$$\begin{aligned} \mathrm{tr}\big(\mathbf{H}(\mathcal{S}) Q_s \mathbf{H}^\dagger(\mathcal{S})\big) &= \sum_{i\in\mathcal{S}} \rho_i \,\mathrm{tr}\big(\mathbf{h}_i \mathbf{h}_i^\dagger\big) & (4.84)\\ &= \sum_{i\in\mathcal{S}} \rho_i \,|\mathbf{h}_i|^2 & (4.85)\\ &\leq \max_{i\in\mathcal{S}} |\mathbf{h}_i|^2 \sum_{i\in\mathcal{S}} \rho_i & (4.86)\\ &\leq \max_{1\leq i\leq n} |\mathbf{h}_i|^2 \sum_{i\in\mathcal{S}} \rho_i, & (4.87) \end{aligned}$$

where ρ_i is the transmit power of the secondary user i. Let

$$P_{sum} = \sum_{i \in \mathcal{S}} \rho_i, \tag{4.88}$$

and

$$h_{max} = \max_{1 \le i \le n} |\mathbf{h}_i|^2. \tag{4.89}$$

We can rewrite the right hand side of (4.83) as

$$R_{mac} \le m \log \big(1 + \frac{1}{m} h_{max} P_{sum}\big). \tag{4.90}$$

We first bound P_{sum} and formulate an optimization as:

$$\begin{aligned} &\max_{\mathcal{S}, \{\rho_i\}} \ P_{sum} \\ \text{subject to: } &\rho_i \le \rho_s \text{ for } i \in \mathcal{S} \\ &\big[\mathbf{G}_p \, Q_s \, \mathbf{G}_p^\dagger\big]_{\ell,\ell} \le \Gamma \text{ for } 1 \le \ell \le N_p, \end{aligned} \tag{4.91}$$

which is a standard linear programming, and the solution is denoted by P_{sum}^*. Then, P_{sum}^* is the maximum total transmit power, depending on the channel realizations for each transmission.

Subject to the interference constraints on the primary, the user selection and power allocation are coupled, and a direct analysis is difficult. Instead, we will find an upper bound for P_{sum}^*. Notice that the total interference (on all primary users) caused by the secondary user i is $\rho_i |\mathbf{g}_{p,i}|^2$, where $\mathbf{g}_{p,i}$ is the vector of channel coefficients from the secondary i to all N_p primary users. We relax the set of individual interference constraints in (4.91) with a single sum interference constraint:

$$\sum_{i \in \mathcal{S}} \rho_i |\mathbf{g}_{p,i}|^2 \le N_p \Gamma. \tag{4.92}$$

Notice that $\mathbf{g}_{p,i}$ corresponds to a certain column in $\mathbf{G}_p$.

Order the cross channel gains $\{|\mathbf{g}_{p,i}|^2\}_{i=1}^n$ of all the secondary users and denote the ordered cross channel gains by

$$|\tilde{\mathbf{g}}_{p,1}|^2 \le |\tilde{\mathbf{g}}_{p,2}|^2 \le \cdots \le |\tilde{\mathbf{g}}_{p,n}|^2. \tag{4.93}$$

Then, we further relax the sum interference constraint (4.92) by replacing $\{|\mathbf{g}_{p,i}|^2\}_{i\in\mathcal{S}}$ with the first $|\mathcal{S}|$ smallest cross channel gains $\{|\tilde{\mathbf{g}}_{p,i}|^2\}_{i=1}^{|\mathcal{S}|}$. Thus, we have:

$$\begin{aligned} &\max_{\mathcal{S},\{\rho_i\}} \; P_{sum} \\ \text{subject to:} &\sum_{i=1}^{|\mathcal{S}|} \rho_i |\tilde{\mathbf{g}}_{p,i}|^2 \le N_p\Gamma \\ &\rho_i \le \rho_s \text{ for } 1 \le i \le |\mathcal{S}|. \end{aligned} \tag{4.94}$$

For any channel realization, the solution for the above problem, denoted by $P^*_{sum,1}$, is always greater than, or equal to P^*_{sum}. Notice that $P^*_{sum,1}$ is also a random variable. Since $\{|\tilde{\mathbf{g}}_{p,i}|^2\}$ is non-decreasing in i, the set of $\{\rho_i\}$ that achieves $P^*_{sum,1}$ satisfies $\rho_i \ge \rho_j$, for $i \le j$. In other words, we have $\rho_i = \rho_s$, for $i = 1$ to $|\mathcal{S}| - 1$, and $\rho_i \le \rho_s$, for $i = |\mathcal{S}|$.

Let S_{max} be the maximum value of $|\mathcal{S}|$ that satisfies the constraint

$$\rho_s \sum_{i=1}^{|\mathcal{S}|-1} |\tilde{\mathbf{g}}_{p,i}|^2 \le N_p\Gamma. \tag{4.95}$$

We have

$$P^*_{sum,1} \le \rho_s S_{max}, \tag{4.96}$$

where in (4.96) we have an inequality because the constraint (4.95) is relaxed by discarding $\rho_{|\mathcal{S}|}$ compared to the interference constraint in (4.94) .

Now, we focus on bounding $\rho_s S_{max}$. For any positive integer k, we have

$$P_r\big(S_{max} < k\big) \ge P_r\big(\sum_{i=1}^{k-1} |\tilde{\mathbf{g}}_{p,i}|^2 > \frac{N_p\Gamma}{\rho_s}\big), \tag{4.97}$$

which comes from the fact that the event of the right hand side implies the event of the left hand side. Notice that $\sum_{i=1}^{k-1} |\tilde{\mathbf{g}}_{p,i}|^2$ is a sum of least order statistics out of $\{|\mathbf{g}_{p,i}|^2\}_{i=1}^n$ with i.i.d. Gamma$(N_p, 1)$ distributions. We apply some results in the development of [42, Proposition 12], and obtain[7]

$$P_r\big(\sum_{i=1}^{f(n)-1} |\tilde{\mathbf{g}}_{p,i}|^2 > \frac{N_p\Gamma}{\rho_s}\big) > 1 - O\big(\frac{1}{f(n)}\big), \tag{4.98}$$

[7] For our case, $\frac{1}{\lambda} = \gamma = N_p$.

where $f(n) = c_0 n^{\frac{1}{N_p+1}}$, and $c_0 = \left(\frac{\Gamma(N_p+1)}{(1-\epsilon)\rho_s} N_p^{-\frac{1}{N_p}}\right)^{\frac{N_p}{N_p+1}}$. For large N_p and small ϵ, $c_0 \approx \frac{\Gamma}{\rho_s}(N_p+1)$.

Let $k = f(n)$ in (4.97) and combine with (4.98):

$$P_r\bigg(\rho_s S_{max} < \rho_s\, f(n)\bigg) > 1 - O\big(n^{-\frac{1}{N_p+1}}\big). \tag{4.99}$$

After characterizing $\rho_s S_{max}$, now we return to P^*_{sum}. To simplify notation, we denote

$$\bar{p}_{sum} = \rho_s\, f(n). \tag{4.100}$$

Because $P^*_{sum} \leq P^*_{sum,1} \leq \rho_s S_{max}$ for any channel realizations, from (4.99), we have

$$\begin{aligned} P_r\bigg(P^*_{sum} \geq \bar{p}_{sum}\bigg) &= 1 - P_r\bigg(P^*_{sum} < \bar{p}_{sum}\bigg) \\ &< 1 - P_r\bigg(\rho_s S_{max} < \bar{p}_{sum}\bigg) \\ &< O\big(n^{-\frac{1}{N_p+1}}\big). \end{aligned} \tag{4.101}$$

Now, we complete the analysis of P^*_{sum}, and move to h_{max}. Because $\{|\mathbf{h}_i|^2\}_{i=1}^n$ have i.i.d. Gamma$(m, 1)$ distributions, using the similar arguments developed in Lemma 4.4.1, we obtain

$$P_r\bigg(h_{max} > \zeta_n\bigg) = O\big(\frac{1}{\log n}\big) \tag{4.102}$$

$$\mathbb{E}\big[h_{max} \,\big|\, h_{max} > \zeta_n\big] < O(n \log n), \tag{4.103}$$

where ζ_n is a deterministic sequence satisfying

$$\zeta_n = \log n + m \log\log n + O(\log\log\log n). \tag{4.104}$$

Now we are ready to develop the upper bound for the secondary throughput. Since $P_{sum} \leq P^*_{sum}$, from (4.90), we have

$$
\begin{aligned}
\mathcal{R}_{mac} &\leq m\mathbb{E}_{\mathbf{H},P}\Big[\log\Big(1+\frac{1}{m}h_{max}P^*_{sum}\Big)\Big] && (4.105)\\
&= m\mathbb{E}_{\mathbf{H},P}\Big[\log\Big(1+\frac{1}{m}h_{max}P^*_{sum}\Big)\Big|P^*_{sum} < \bar{p}_{sum}\Big] \times P_r\big(P^*_{sum} < \bar{p}_{sum}\big)\\
&\quad + m\mathbb{E}_{\mathbf{H},P}\Big[\log\Big(1+\frac{1}{m}h_{max}P^*_{sum}\Big)\Big|P^*_{sum} \geq \bar{p}_{sum}\Big] \times P_r\big(P^*_{sum} \geq \bar{p}_{sum}\big) && (4.106)\\
&\leq m\mathbb{E}_{\mathbf{H}}\Big[\log\Big(1+\frac{1}{m}h_{max}\bar{p}_{sum}\Big)\Big]\cdot 1\\
&\quad + m\mathbb{E}_{\mathbf{H}}\Big[\log\Big(1+\frac{1}{m}h_{max}\rho_s n\Big)\Big]\cdot O\big(n^{-\frac{1}{N_p+1}}\big) && (4.107)\\
&\leq m\mathbb{E}_{\mathbf{H}}\Big[\log\Big(1+\frac{1}{m}h_{max}\bar{p}_{sum}\Big)\Big|h_{max} \leq \zeta_n\Big] \times P_r\big(h_{max} \leq \zeta_n\big)\\
&\quad + m\mathbb{E}_{\mathbf{H}}\Big[\log\Big(1+\frac{1}{m}h_{max}\bar{p}_{sum}\Big)\Big|h_{max} > \zeta_n\Big] \times P_r\big(h_{max} > \zeta_n\big)\\
&\quad + m\mathbb{E}_{\mathbf{H}}\Big[\log\Big(1+\frac{1}{m}h_{max}\rho_s n\Big)\Big|h_{max} \leq \zeta_n\Big] \times P_r\big(h_{max} \leq \zeta_n\big)O\big(n^{-\frac{1}{N_p+1}}\big)\\
&\quad + m\mathbb{E}_{\mathbf{H}}\Big[\log\Big(1+\frac{1}{m}h_{max}\rho_s n\Big)\Big|h_{max} > \zeta_n\Big]P_r\big(h_{max} > \zeta_n\big)O\big(n^{-\frac{1}{N_p+1}}\big) && (4.108)\\
&\leq m\log\Big(1+\frac{1}{m}\,\zeta_n\,\bar{p}_{sum}\Big)\cdot 1\\
&\quad + m\log\Big(1+\frac{\bar{p}_{sum}}{m}\,\mathbb{E}\big[h_{max}\,\big|\,h_{max} > \zeta_n\big]\Big) \times P_r\big(h_{max} > \zeta_n\big)\\
&\quad + m\log\Big(1+\frac{1}{m}\zeta_n\,\rho_s n\Big)\cdot 1\cdot O\big(n^{-\frac{1}{N_p+1}}\big)\\
&\quad + m\log\Big(1+\frac{\rho_s n}{m}\mathbb{E}\big[h_{max}\,\big|\,h_{max} > \zeta_n\big]\Big) \times P_r\big(h_{max} > \zeta_n\big)O\big(n^{-\frac{1}{N_p+1}}\big) && (4.109)\\
&\leq m\log\Big(1+\frac{1}{m}\,\zeta_n\,\bar{p}_{sum}\Big) + m\log\Big(1+\frac{\bar{p}_{sum}}{m}O(n\log n)\Big)O(\frac{1}{\log n})\\
&\quad + m\log\Big(1+\frac{1}{m}\zeta_n\rho_s n\Big)\,O\big(n^{-\frac{1}{N_p+1}}\big)\\
&\quad + m\log\Big(1+\frac{\rho_s n}{m}O(n\log n)\Big)O(\frac{1}{\log n})O\big(n^{-\frac{1}{N_p+1}}\big), && (4.110)
\end{aligned}
$$

where the second term in (4.107) comes from using (4.101) as well as the fact that P^*_{sum} is upper bounded by $\rho_s n$. In (4.109), we apply the Jensen inequality to obtain the second and fourth terms. Using (4.102) and (4.103), we have the second and fourth terms in (4.110). Finally, by substituting $\bar{p}_{sum}$ and ζ_n, we obtain

$$\mathcal{R}_{mac} \leq \frac{m}{N_p+1}\log n + O(\log\log n). \tag{4.111}$$

This concludes the proof of this theorem.

4.7.3 Proof of Lemma 4.4.1

First, we prove (4.45). Let $Z = |\mathbf{h}_i^\dagger \phi_j|^2$ and $Y = \theta\big(\sum_{k\neq j}|\mathbf{h}_i^\dagger \phi_j|^2 + |\mathbf{g}_{s,i}|^2\big)$. Then, Z has the exponential distribution, and Y has the Gamma$\big((m+M_p-1),\theta\big)$ distribution. We can write

$$L_i = \frac{Z}{c+Y}, \tag{4.112}$$

where $c = \frac{m}{\rho}$. Conditioned on Y, the pdf of L_i is given by

$$\begin{aligned} f_L(x) &= \int_0^\infty f_{L|Y}(x|y) f_Y(y) dy && (4.113)\\ &= \int_0^\infty (c+y)e^{-(c+y)x} \times \frac{y^{m+M_p-1}e^{-y/\theta}}{(m+M_p-1)!\,\theta^{m+M_p}} dy && (4.114)\\ &= \frac{e^{-cx}}{(1+\theta x)^{m+M_p}}\big(c(1+\theta x)+\theta(m+M_p-1)\big). && (4.115)\end{aligned}$$

So the cdf of L_i is

$$\begin{aligned} F_L(x) &= 1 - \int_x^\infty f_L(t)dt && (4.116)\\ &= 1 - \frac{e^{-cx}}{(1+\theta x)^{m+M_p-1}}. && (4.117)\end{aligned}$$

We define a growth function as

$$\begin{aligned} g_L(x) &= \frac{1-F_L(x)}{f_L(x)} && (4.118)\\ &= \frac{1+\theta x}{c(1+\theta x)+\theta(m+M_p-1)}. && (4.119)\end{aligned}$$

Since $\lim_{x\to\infty} g_L'(x) = 0$, the limiting distribution of $L_{max} = \max_{1\leq i\leq n} L_i$ exists [49]:

$$\lim_{n\to\infty}\big(F_L(b_n+a_n x)\big)^n = e^{-e^{-x}}, \tag{4.120}$$

where $b_n = F_L^{-1}(1-1/n)$ and $a_n = g_L(b_n)$. In general, an exact closed-form solution for a_n and b_n is intractable, but an approximation can be obtained, which is sufficient for asymptotic analysis. After manipulating (4.117), we have

$$b_n = \frac{1}{c}\log n - \frac{m+M_p-1}{c}\log\log n + O\big(\log\log\log n\big), \tag{4.121}$$

and thus

$$a_n = \frac{1}{c} + O\big(\frac{1}{\log n}\big). \tag{4.122}$$

It is straightforward to verify $\lim_{n\to\infty}\big(ng_L'(b_n)\big) = \infty$, so we apply the expansion developed in [50, Eq. (22)]

$$\big(F_L(b_n + a_n x)\big)^n = \exp\Bigg(-\exp(-x + \Theta(\frac{x^2}{\log^2 n}))\Bigg). \tag{4.123}$$

Let $x = -\log\log n$ in (4.123) we obtain (4.45).

Now, we prove (4.46) and (4.47). Since U_i is similar to L_i, except that the denominator now has the Gamma$\big(M_p, \theta\big)$ distribution. Following the same steps of obtaining (4.123), we have the expansion of the cdf of U_{max}:

$$\big(F_U(d_n + c_n x)\big)^n = \exp\Bigg(-\exp(-x + \Theta(\frac{x^2}{\log^2 n}))\Bigg), \tag{4.124}$$

where

$$d_n = \frac{1}{c}\log n - \frac{M_p}{c}\log\log n + O\big(\log\log\log n\big), \tag{4.125}$$

and

$$c_n = \frac{1}{c} + O\big(\frac{1}{\log n}\big), \tag{4.126}$$

where (4.46) follows by substituting $x = \log\log n$ into (4.124).

Finally, because $\mathbb{E}[U_{max}] < n\mathbb{E}[U_i]$ [49], we have

$$\begin{aligned}\mathbb{E}\Bigg[U_{max}\,\Bigg|\,U_{max} > d_n + \frac{1}{c}\log\log n\Bigg] &\le \frac{n\mathbb{E}[U_i]}{P_r\big(U_{max} > d_n + \frac{1}{c}\log\log n\big)} \\ &= \Theta(n\log n),\end{aligned} \tag{4.127}$$

where we use (4.46) in the last equality.

4.7.4 Proof of Theorem 4.4.2

We first find a lower bound for the secondary throughput $\mathcal{R}_{bc}$. Conditioned on $P = \rho$, the throughput is denoted $\mathcal{R}_{bc|P}(\rho)$. Let $l_n = b_n - \frac{\rho}{m} \log\log n$, where b_n is given by Lemma 4.4.1. Using (4.44), the conditional throughput $\mathcal{R}_{bc|P}(\rho)$ can be bounded as

$$\begin{aligned}
\mathcal{R}_{bc|P}(\rho) &\geq m\mathbb{E}\Big[\log\big(1 + L_{max}\big) \,\Big|\, P = \rho \Big] &(4.128)\\
&\geq m\mathbb{E}\Big[\log\big(1 + L_{max}\big) \,\Big|\, L_{max} \geq l_n,\, P = \rho \Big] \times P_r\big(L_{max} \geq l_n \,\big|\, P = \rho\big) &(4.129)\\
&> m\bigg(\log\big(\frac{\rho}{m} \log n\big) + O\big(\frac{\log\log n}{\log n}\big) \bigg) \times \Big(1 - \Theta\big(n^{-1}\big) \Big) &(4.130)\\
&= m \log\big(\frac{\rho}{m} \log n\big) + O\big(\frac{\log\log n}{\log n}\big). &(4.131)
\end{aligned}$$

From (4.128) to (4.129), we discard the non-negative term associated with the event $\{L_{max} < l_n\}$. Using (4.45) from Lemma 4.4.1 and the identity $\log(x + y) = \log x + \log(1 + y/x)$, for $x, y > 0$, we have (4.130).

Now we take the expectation with respect to P. From (4.37), we have

$$P > \frac{m\Gamma}{\max_{1 \leq i \leq N_p} |\mathbf{g}_{p,i}^{\dagger}|^2 + m\Gamma/P_s}, \tag{4.132}$$

where $\mathbf{g}_{p,i}^{\dagger}$ is the $1 \times m$ vector of channel coefficients from the secondary base station to the primary user i. Let the pdf of $\max_{1 \leq i \leq N_p} |\mathbf{g}_p(i)|^2$ be $f_{g_p}(x)$. Because (4.132) holds for any channel realization, we have

$$\begin{aligned}
\mathcal{R}_{bc} &> \int_0^{\infty} m \log\left(\frac{\Gamma \log n}{x + m\Gamma/P_s} \right) f_{g_p}(x)\, dx + O\bigg(\frac{\log\log n}{\log n} \bigg) &(4.133)\\
&\geq m \log\left(\frac{\Gamma \log n}{\tilde{\mu}_1 + m\Gamma/P_s} \right) + O\bigg(\frac{\log\log n}{\log n} \bigg) &(4.134)\\
&= m \log\big(\Gamma \log n\big) - m \log\big(\tilde{\mu}_1 + m\Gamma/P_s\big) + O\bigg(\frac{\log\log n}{\log n} \bigg), &(4.135)
\end{aligned}$$

where (4.134) comes from the convexity of $\log(a + \frac{b}{x+c})$ and

$$\tilde{\mu}_1 = \mathbb{E}[\max_{1 \leq i \leq N_p} |\mathbf{g}_p(i)|^2]. \tag{4.136}$$

To find an upper bound, we still begin with the conditional throughput $\mathcal{R}_{bc|P}(\rho)$. Let $u_n = d_n + \frac{\rho}{m}\log\log n$, where d_n is given by Lemma 4.4.1. Then

$$\begin{aligned}
\mathcal{R}_{bc|P}(\rho) &\leq m\mathbb{E}\Big[\log\big(1+U_{max}\big)\,\Big|\,P=\rho\Big] && (4.137)\\
&\leq m\mathbb{E}\Big[\log\big(1+U_{max}\big)\,\Big|\,U_{max}<u_n,\ P=\rho\Big]\times P_r\big(U_{max}<u_n\big|P=\rho\big) && (4.138)\\
&\quad + m\mathbb{E}\Big[\log\big(1+U_{max}\big)\,\Big|\,U_{max}\geq u_n,\ P=\rho\Big]\times P_r\big(U_{max}\geq u_n\big|P=\rho\big) && (4.139)\\
&< m\log(1+u_n)\big(1-\Theta\big(\frac{1}{\log n}\big)\big)\\
&\quad + m\log\big(1+\mathbb{E}[U_{max}\,|\,U_{max}\geq u_n,\ P=\rho]\big)\times\Theta\big(\frac{1}{\log n}\big) && (4.140)\\
&< m\log(1+\frac{\rho}{m}\log n)+O(1), && (4.141)
\end{aligned}$$

where (4.137) comes from (4.44). We apply (4.46) in Lemma 4.4.1 and the Jensen inequality to obtain (4.140). Using (4.47) in Lemma 4.4.1 and substituting u_n, we obtain (4.141).

After calculating an upper bound for the conditional throughput, we average over P. From (4.37), we have

$$P \leq \frac{m\Gamma}{\max_{1\leq i\leq N_p}|\mathbf{g}_{p,i}^{\dagger}|^2}. \tag{4.142}$$

We denote

$$\frac{1}{\tilde{\mu}_2} = \mathbb{E}\big[1/\max_{1\leq i\leq N_p}|\mathbf{g}_{p,i}^{\dagger}|^2\big]. \tag{4.143}$$

Then, by the Jensen inequality, we have

$$\begin{aligned}
\mathcal{R}_{bc} &< m\log\big(1+\frac{\log n}{m}\mathbb{E}[P]\big)+O(1) && (4.144)\\
&< m\log\big(1+\frac{\Gamma}{\tilde{\mu}_2}\log n\big)+O(1) && (4.145)\\
&= m\log(\Gamma\log n)-m\log\tilde{\mu}_2+O(1), && (4.146)
\end{aligned}$$

where (4.145) holds since $\mathbb{E}[P]\leq\frac{m\Gamma}{\tilde{\mu}_2}$. The theorem follows.

CHAPTER 5
HYBRID OPPORTUNISTIC SCHEDULING IN COGNITIVE RADIO NETWORKS

5.1 Introduction

This chapter studies an underlay cognitive multiple-access (MAC) channel with n transmitters, in the presence of a primary system with M_p transmitters and N_p receivers. The primary and secondary systems are subject to mutual interference, where the secondary must comply with a set of interference power constraints imposed by the primary. The objective is to design a user scheduling method that exploits multiuser diversity in both cross links and secondary links, so that the secondary sum-rate (throughput) is maximized, while the interference induced on the primary is strictly bounded.

A brief overview of the past work is as follows. Zhang et al. [37] studied the power allocation of a secondary system under various power and interference constraints. Multiple antennas at the secondary transmitter were exploited by [36] to balance the secondary throughput and the interference on the primary. Recently, ideas from opportunistic communication [39] have been applied in underlay cognitive radios. Tajer et al. [51] analyzed a parallel cognitive network and found a growth rate of $\Theta(\log\log n)$ for the throughput. The throughput limits of cognitive broadcast and MAC channel were analyzed [52, 53] (Chapter 4), where [52] randomly activates multiple secondary transmitters with interference smaller than a threshold. Jamal et al. [41] and Shen et al. [54] found that the secondary throughput can be increased by simultaneously activating as many secondary transmitters as possible. The multiuser diversity gain in cognitive networks was also studied by Hong et al. [55], Zhang et al. [56] and Ban et al. [57], showing that by selecting the secondary user with the highest signal-to-interference-and-noise ratio (SINR) under the primary interference constraints, the secondary throughput can grow as $\Theta(\log\log n)$.

The main results of this chapter are as follows.

- We propose a two-step (hybrid) opportunistic scheduling that pre-selects a set of secondary transmitters with small interference, and from among them activates multiple transmitters with large secondary-channel gain. The pre-selection step provides cross-link diversity to minimize interference, while the second step provides multi-user diversity to improve the secondary throughput. The result is a throughput growing as $\Theta(\log n)$, which improves on the growth rate of $\Theta(\log\log n)$ in [55–57]. Furthermore, a 20-30% throughput gain is obtained compared with Chapter 4 for up to 200 secondary users. The proposed scheduling method is shown to be optimal asymptotically, and can reduce the interference on the primary proportionally to n^{-q}, while the secondary throughput grows proportionally to $\frac{1-qN_p}{N_p+1}\log n$, for $0 \le q \le \frac{1}{N_p}$.

- We characterize the (asymptotically) optimal number of active secondary transmitters as a function of the primary interference constraint, the secondary transmit power and n. To achieve the asymptotically optimal secondary throughput, the number of active transmitters must be proportional to $n^{\frac{1}{N_p+1}}$.

- The issue of fairness is studied; this issue arises when the node channel statistics are not identical. A method is proposed to ensure user fairness and the effect of a fairness constraint on asymptotic throughput is analyzed. It is shown that the modified scheduling method achieves the same optimal growth rate for the throughput, i.e., the fairness constraint does not affect the growth rate of the throughput for this algorithm.

The following asymptotic notations are used in this chapter. For sufficiently large n,

$$\begin{array}{lll}
f(n) = O\big(g(n)\big): & \exists c_1 & |f(n)| < c_1|g(n)| \\
f(n) = \Theta\big(g(n)\big): & \exists c_1, c_2 & c_2|g(n)| < |f(n)| < c_1|g(n)| \\
f(n) = o\big(g(n)\big): & \forall \epsilon > 0 & |f(n)| < \epsilon|g(n)|
\end{array}$$

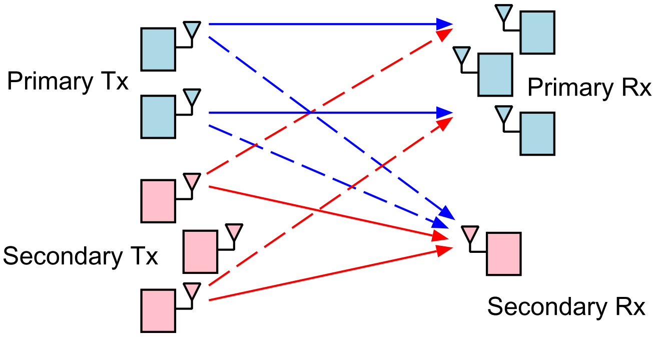

Figure 5.1. Multiple access cognitive radios.

5.2 System Model

We consider a multiple-access (MAC) secondary system that coexists with a primary system, as shown in Figure 5.1. The primary system consists of M_p transmitters and N_p receivers,[1] where each transmitter communicates with one or more receivers, and vice versa. The primary and secondary are subject to mutual interference from each other which is treated as noise. The interference from the secondary to each primary receiver must be smaller than a pre-defined interference temperature (threshold). For simplicity, all nodes are assumed to be single-antenna.

A block-fading channel model is assumed where all channel coefficients are independent, identically distributed (i.i.d.) circularly-symmetric complex Gaussian with zero mean and unit variance, denoted by $\mathcal{CN}(0,1)$. For each transmission, a subset of secondary transmitters are activated; the collection of selected (active) transmitters is denoted by $\mathcal{S}$. The signal at the secondary receiver is:

$$y = \sum_{i \in \mathcal{S}} \sqrt{P_i}\, h_i\, x_i + \sum_{\ell=1}^{M_p} \sqrt{P_p}\, g_{s,\ell}\, x_{p,\ell} + w, \tag{5.1}$$

where h_i is the channel coefficient from the secondary transmitter i to the secondary receiver. The secondary transmitter i sends a signal x_i with power P_i, which is subject to a short term power constraint, i.e., $P_i \leq P$ for $1 \leq i \leq n$. The cross-channel coefficient from the primary

[1] In this chapter, M_p and N_p are assumed to be bounded, i.e., not scaling with n.

transmitter ℓ to the secondary receiver is $g_{s,\ell}$. The primary transmitter ℓ sends a signal $x_{p,\ell}$ with power P_p for $1 \leq \ell \leq M_p$. The additive noise w has the distribution $\mathcal{CN}(0,1)$.

The interference power (caused by the secondary transmitters) on the primary receiver j is

$$I_j = \sum_{i \in \mathcal{S}} P_i |g_{ji}|^2, \tag{5.2}$$

where g_{ji} is the cross-channel coefficient from the secondary transmitter i to the primary receiver j. For clarity of exposition, all the primary receivers are assumed to tolerate a short-term interference power Γ; the case of unequal tolerances can be studied similarly (see Remark 5.3.1). We have

$$I_j \leq \Gamma, \quad \text{for} \quad 1 \leq j \leq N_p. \tag{5.3}$$

Throughout this chapter, we assume the secondary receiver knows the secondary-channel coefficients $\{h_i\}$ but does not know any other channels (see Remark 5.3.2 for more details). We refer to the secondary forward channel simply as the *secondary-channel*, and the secondary cross-channel to the primary receiver as the *cross-channel*.

5.3 Scheduling in Cognitive MAC channel

A scheduling scheme determines a set of active secondary transmitters $\mathcal{S}$ and their power $\{P_i\}_{i \in \mathcal{S}}$. The corresponding average secondary sum-rate (throughput) is given by

$$R_{mac} = \mathbb{E}\left[\log\left(1 + \frac{G_{sum}}{1 + I_p}\right)\right], \tag{5.4}$$

where

$$G_{sum} = \sum_{i \in \mathcal{S}} P_i |h_i|^2, \quad I_p = P_p \sum_{\ell=1}^{M_p} |g_{s,\ell}|^2. \tag{5.5}$$

The statistics of G_{sum} depends on the associated scheduling rule, and are independent of I_p, the interference from the primary.

5.3.1 Hybrid Opportunistic Scheduling

A secondary user scheduling should maximize the (average) secondary throughput, while satisfying the primary-imposed interference constraints. However, such two objectives often conflict. To increase the throughput, we want to activate many transmitters with large secondary-channel gains, but these transmissions may violate the interference constraints. Since the interference from various concurrent transmissions will add up, the scheduling of secondary transmitters is interdependent. We may choose many transmitters operating at low power, or a few transmitters at high power. Moreover, even for a fixed number of transmitters, reducing power from one transmitter allows increasing power from other transmitters. In general, the search for the optimal transmitter set and transmit power is a variation of the knapsack problem, which is NP-complete. To simplify the problem, we adopt a decoupling power policy that is shown to be asymptotically optimal later on. This is an on-off power policy where each transmitter *either operates at maximum power P or remains silent.* Then, the scheduling scheme is as follows:

Selection of Eligible Transmitters

The scheduling process has two parts. In its first part, we concentrate on limiting the interference, thus favoring transmitters with small cross-channel gains. Specifically, we only allow transmitters that do not violate an interference quota α on *each* primary receiver. The collection of such transmitters is defined as the eligible transmitter set:

$$\mathcal{A} = \Bigg\{ i : P|g_{ji}|^2 < \alpha, \, \forall \, 1 \leq j \leq N_p \Bigg\}. \tag{5.6}$$

This step can be considered as *opportunistic interference avoidance.* Recall that each primary receiver can tolerate interference power Γ. Once the maximum interference generated by each secondary transmitter is capped, the total interference at each primary receiver is guaranteed to be tolerable if no more than $k_s = \frac{\Gamma}{\alpha}$ eligible secondary transmitters are in operation.[2]

[2] For the purposes of analysis α is allowed to take any small and positive value, but for practical purposes it can be limited to the values that make k_s to be an integer.

Selection of Active Transmitters

Now we choose from among the eligible transmitters those who will actually transmit. Up to k_s secondary transmitters will be chosen that have high secondary-channel gains (SNRs), therefore producing multiuser diversity. The ordered channel gains of eligible transmitters are denoted by:

$$|\tilde{h}_1|^2 \geq |\tilde{h}_2|^2 \geq \cdots \geq |\tilde{h}_M|^2, \tag{5.7}$$

where $|\tilde{h}_i|^2$ is the ith largest channel gain of transmitters in $\mathcal{A}$, and $M = |\mathcal{A}|$ is the size of $\mathcal{A}$. Note that M is a random variable. If $M > k_s$, the first k_s transmitters in the above order will be active simultaneously. If $M \leq k_s$, then all the M eligible transmitters will operate.

The above two-step scheme is called *Hybrid Opportunistic Scheduling* in the sense that it is driven by a hybrid of two criteria: Minimizing interference and maximizing throughput. This selection process requires neither exhaustive search nor joint power control among secondary transmitters, but it still guarantees compliance with the pre-defined interference threshold and captures the multiuser diversity gain. In addition, this scheduling is simple to design; the only parameter to consider is the interference quota α (thus k_s), which will be studied in the sequel.

Remark 5.3.1 *Hybrid Opportunistic Scheduling still applies when primary receivers tolerate unequal amounts of interference, e.g., Γ_j for $1 \leq j \leq N_p$. In this case, we design a separate interference quota for* each *primary receiver, i.e., $\alpha_j = \frac{\Gamma_j}{k_s}$, and re-define the eligible transmitter set as*

$$\mathcal{A}_{neq} = \left\{ i : P|g_{ji}|^2 < \alpha_j, \forall\, 1 \leq j \leq N_p \right\}, \tag{5.8}$$

such that the transmission of any k_s eligible secondary transmitters complies with all the interference constraints. Notice that the selection of active transmitters is unaffected. One can show that most of the analysis and results in this chapter still follow in a similar manner.

Remark 5.3.2 *We briefly discuss the CSI requirement of the proposed scheme. First, each secondary transmitter compares its cross-channel gains*[3] *to a threshold to evaluate its eligibility. Then,* only *among eligible transmitters each sends* 1*-bit to inform the secondary receiver. The secondary-channel of eligible transmitters can be directly estimated at the receiver side. Therefore, this scheduling method requires little exchange of CSI. The thresholding operation of our method is essentially a distributed decision making process that significantly reduces the CSI overhead compared with methods that choose the least interfering secondary [41], because ranking is by necessity a centralized process and requires all nodes to communicate their cross-link to the receiver.*

5.3.2 Throughput Analysis

Now we study the throughput achieved by the proposed Hybrid Opportunistic Scheduling. We first derive the average secondary throughput, and then maximize the throughput over α. Under the proposed scheduling, we have

$$G_{sum} = P \sum_{i=1}^{\min(k_s,M)} |\tilde{h}_i|^2, \tag{5.9}$$

which involves a sum of order statistics whose properties are given by the following lemma.

Lemma 5.3.1 *Let a and b be large positive integers with $b \geq a$, and $S_b^a(\rho)$ be the sum of the highest a order statistics out of b i.i.d. exponentials with mean ρ. For any $0 < \epsilon < 1$,*

$$\mathbb{P}\Big(\big| S_b^a(\rho) - \rho\mu_b^a \big| < \epsilon\rho\mu_b^a \Big) > 1 - O\bigg(\frac{1}{\big(\log b\big)^2} \bigg),$$

$$\mathbb{E}[S_b^a(\rho)] = \rho\mu_b^a,$$

where $\mu_b^a = a \log \frac{b}{a} + a + O(1)$.

[3] The primary receiver emits packets for, e.g., handshake or ACK/NACK, which can be overheard by the secondary transmitter and used for cross-channel gain estimation in a TDD system. Also, under the spectrum leasing model [58], the primary receivers can be expected to actively promote spectrum reuse by transmitting pilots that can be used for cross-channel gain estimation. The latter model applies to both TDD and FDD.

Proof See Section 5.6.1.

Remark 5.3.3 *In Lemma 5.3.1, μ_b^a can be considered as the multiuser diversity gain achieved by selecting the best a out of b users in i.i.d. Rayleigh fading channels. For $a = 1$, it reduces to the case where one transmitter with the highest channel gain is selected, and we have $\mu_b^1 \approx \log b$, a well known result [39, 49]. For $a = b$ (no selection), $S_b^a(\rho)$ obeys Gamma(b, ρ) distribution, and $\rho\mu_b^b \approx \rho b$.*

Based on Lemma 5.3.1 and recalling that $k_s = \frac{\Gamma}{\alpha}$, for sufficiently small α (large k_s), we have the following results.

Theorem 5.3.2 *Consider a secondary MAC with n transmitters, each with power P. This MAC coexists with a primary system with N_p receivers and M_p transmitters each with power P_p. If each primary receiver tolerates interference power Γ, then the average secondary throughput R_{mac} satisfies*

$$R_{mac} \geq \log \frac{\big(\log n - (N_p + 1)\log k_s + N_p \log(\Gamma/P) + 1\big)}{1 + M_p P_p} + \log \frac{P k_s}{1 + M_p P_p} + O\big(\frac{1}{\log n}\big), \tag{5.10}$$

$$R_{mac} \leq \log \frac{\big(\log n - (N_p + 1)\log k_s + N_p \log(\Gamma/P) + 1\big)}{1 + M_p P_p} + \log \frac{P k_s}{1 + M_p P_p} + C_0 + O\big(\frac{1}{\log n}\big), \tag{5.11}$$

for sufficiently large n and k_s, where $C_0 = \log\big(\mathbb{E}[1/(1 + I_p)]\mathbb{E}[1 + I_p]\big)$.

Remark 5.3.4 *The lower bound (5.10) has only a constant gap C_0 relative to the upper bound (5.11) for large n, therefore, for given k_s, R_{mac} scales as $\log\log n$, similar to the results in [55–57]. To achieve this secondary rate, multiuser decoding is required at the secondary receiver for $k_s > 1$, which is unlike TDMA scheduling ($k_s = 1$) where single-user detection is sufficient. Finally, we note that C_0 depends only on the statistics of I_p, the interference from the primary to the secondary (see (5.5)).*

Now, we design the interference quota α (equivalently k_s) to maximize the secondary throughput. Unlike conventional MAC where $k_s = n$ maximizes the sum throughput, in spectrum-sharing networks k_s (thus α) must be carefully designed due to the additional primary interference constraints. If α is very small, the number of eligible transmitters is also small on average, which reduces the multiuser diversity gain achieved by selecting from among the eligible transmitters. If α is very large, $\frac{\Gamma}{\alpha}$ will be small and few transmitters can be activated, thus once again the overall throughput will suffer. Therefore, it is desirable to optimize α (thus k_s), as shown by the following lemma.

Lemma 5.3.3 *For sufficiently large n the optimal number of active secondary transmitters k_s^{opt} satisfies*

$$\left|\frac{k_s^{opt}}{k_s^*} - 1\right| \leq \sqrt{1-\xi},$$

where $k_s^* = \left(\frac{\Gamma}{Pe}\right)^{\frac{N_p}{N_p+1}} n^{\frac{1}{N_p+1}}$ and ξ is given by (5.52).

Proof See Section 5.6.3.

Lemma 5.3.3 asymptotically bounds the optimal number of active secondary transmitters as a function of Γ, P and n. It shows that, essentially, k_s^{opt} cannot be too far from k_s^*. Motivated by this lemma, we choose $k_s = k_s^*$ and in the following theorem obtain a throughput growth rate that is later shown to be asymptotically optimal (See Chapter 5.3.3).

Theorem 5.3.4 *Consider a secondary MAC with n transmitters each with power P. This MAC coexists with a primary system with N_p receivers and M_p transmitters with power P_p. If each primary receiver tolerates interference power Γ, then the average secondary throughput R_{mac} satisfies*

$$R_{mac} \geq \frac{1}{N_p+1}\log n + C_1 + O\left(\frac{1}{\log n}\right), \tag{5.12}$$

$$R_{mac} \leq \frac{1}{N_p+1}\log n + C_1 + C_0 + O\left(\frac{1}{\log n}\right), \tag{5.13}$$

for sufficiently large n by activating k_s^* transmitters, where $C_1 = \frac{N_p}{N_p+1}\log\frac{\Gamma}{Pe} + \log\frac{(N_p+1)P}{1+M_pP_p}$.

Proof Notice that the proof of Theorem 5.3.2 holds for $k_s = \Theta(n^{1/(N_p+1)})$. The theorem follows by substituting k_s^* into (5.10) and (5.11), respectively.

The implications of Theorem 5.3.4 are as follows. Intuitively, the secondary throughput is reduced when the number of primary receivers (constraints) increases. Theorem 5.3.4 explicitly quantifies this: $R_{mac} = \frac{1}{N_p+1}\log n + O(1)$. For small N_p, Hybrid Opportunistic Scheduling achieves a (significant) fraction of the throughput of an ordinary MAC, if n is large enough. The achieved throughput is proven to be optimal asymptotically (with n) in the sequel.

So far we have shown that the multiuser nature of a secondary system can improve the secondary throughput. In fact, this multiuser flexibility can also be used to mitigate the interference on the primary. A tradeoff exists between the primary interference reduction and the secondary throughput enhancement under Hybrid Opportunistic Scheduling, which is described as follows.

Corollary 5.3.5 *Consider the allowable interference on each primary receiver being bounded as $\Theta(n^{-q})$. Then, the average secondary throughput satisfies*

$$R_{mac} = \frac{1 - qN_p}{N_p + 1}\log n + O(1), \tag{5.14}$$

for sufficiently large n under Hybrid Opportunistic Scheduling, where $0 \leq q \leq \frac{1}{N_p}$.

Proof Notice that Theorem 5.3.4 holds for $\Gamma = \Theta(n^{-q})$. The Corollary follows by substituting Γ into the lower and upper bounds given by (5.13).

Based on Corollary 5.3.5, as n increases, Hybrid Opportunistic Scheduling can mitigate interference (to zero) on the primary receivers, while the secondary throughput grows as $\Theta(\log n)$. The allowable interference Γ is made to decline as $\Theta(n^{-q})$, which leads R_{mac} to decrease linearly in q. If Γ is reduced more slowly, e.g., decreasing as $\Theta(\frac{1}{\log n})$, the secondary throughput can increase at a rate of $\frac{1}{N_p+1}\log n$. If we try to mitigate the primary interference faster than $\Theta(n^{-q})$, i.e., $q \geq \frac{1}{N_p}$, the secondary throughput only grows as $o(\log n)$. Therefore, as N_p increases, not only the throughput of the secondary decreases, but also its ability of reducing the interference on the primary.

Remark 5.3.5 *The key to the secondary growth rate* $\Theta(\log n)$ *is to activate multiple secondary transmitters while limiting the interference. This approach is in contrast with [55–57] where a single transmitter with the highest SNR was activated. The main questions to be answered in this work have been: how many secondary transmitters we should activate, how to choose the active secondary transmitters in a relatively straight forward fashion, and how much power should the active transmitters emit to achieve the growth rate while satisfying the interference constraint.*

5.3.3 Optimality of Hybrid Opportunistic Scheduling

We first find an upper bound for the average secondary throughput that applies regardless of transmission strategies.

Theorem 5.3.6 *Consider the coexistence of a secondary MAC with* n *transmitters and a primary system with* N_p *receivers. The* maximum *average throughput of the secondary,* R_{mac}^{opt}, *satisfies*

$$R_{mac}^{opt} \leq \frac{1}{N_p + 1} \log n + O(\log \log n). \tag{5.15}$$

Proof See Section 5.6.4.

The gap between the above upper bound and the throughput attained by Hybrid Opportunistic Scheduling (shown in Theorem 5.3.4) is only on the order of $O(\log \log n)$. This gap is negligible relative to $\Theta(\log n)$ for sufficiently large n, therefore, Hybrid Opportunistic Scheduling asymptotically attains the maximum throughput:

$$\lim_{n \to \infty} \frac{R_{mac}}{R_{mac}^{opt}} = 1. \tag{5.16}$$

Remark 5.3.6 *The growth rate $\Theta(\log n)$ can also be attained by activating secondary users simply according to the least interference, i.e., only based on cross-channel gains [42,52]. The similarity of growth rates may tempt one to say that there is no gain in utilizing secondary channel information [42]. However, similarity of growth rates hides $o(\log n)$ throughput gains by the two-step (hybrid) approach that are highly nontrivial and practically important. For instance, our results show throughput gains of around 20-30% over Chapter 4 (see Figure 5.3) by selecting the users with large secondary-channel gain.*

5.4 Scheduling under non-I.I.D. Link Statistics

In this section, we consider a network where neither the secondary-channels nor cross-channels are identically distributed. This is a practical scenario due to, e.g., different path losses for various links. Assuming that the channel gains obey one out of a finite number of distributions, we enumerate them with the variable $d \in \{1, \ldots, D\}$. Specifically, each user has a secondary-channel gain and cross-channel gain that obeys the exponential distribution with parameter ρ_d and λ_d, respectively. The number of users in each of these groups is $\beta_d n$, where $\sum_{d=1}^{D} \beta_d = 1$.

The secondary transmitters that enjoy larger ρ_d and smaller λ_d have a higher probability to be active under Hybrid Opportunistic Scheduling, so user fairness is no longer guaranteed. In the following, we extend Hybrid Opportunistic Scheduling to ensure a (long-term) temporal fairness [59, 60] in the sense that each secondary transmitter has equal probability (time fraction) to be active. For clarity of exposition, we consider $M_p = N_p = 1$, i.e., one pair of primary transmitter and receiver.

Our strategy is to design the interference quota for Group d to be proportional to λ_d, such that all transmitters have equal eligible probability. More precisely, the interference quota for Group d is

$$\alpha_d = \frac{\lambda_d \Gamma}{k_s \sum_{j=1}^{D} \beta_j \lambda_j}, \quad \text{for } 1 \leq d \leq D. \tag{5.17}$$

The corresponding eligible transmitter set for Group d is

$$\mathcal{A}_d = \Big\{ i : P|g_{ji}|^2 < \alpha_d, \, \forall\, 1 \leq j \leq N_p, \, i \in \text{Group } d \Big\}.$$

Therefore, the eligible probability of *any* transmitter is

$$p' \approx \frac{\Gamma}{k_s P \sum_{d=1}^{D} \beta_j \lambda_j}. \tag{5.18}$$

Then, we *separately* select the best (up to) eligible $\beta_d\, k_s$ transmitters from among *each* group. One can verify that the above modifications ensure both the fairness requirement and the primary interference restriction. We have the following lemma:

Lemma 5.4.1 *For the network described above, the average secondary throughput R_{mac} satisfies*

$$R_{mac} \geq \log \frac{Pk_s\big(\log \frac{n\,p'}{k_s} + 1\big)}{1+P_p} + \log \sum_{d=1}^{D} \rho_d \beta_d + O(\frac{1}{\log n}),$$
$$R_{mac} \leq \log \frac{Pk_s\big(\log \frac{n\,p'}{k_s} + 1\big)}{1+P_p} + \log \sum_{d=1}^{D} \rho_d \beta_d + C_0 + O(\frac{1}{\log n}),$$

for sufficiently large n under the modified Hybrid Opportunistic Scheduling.

Proof For brevity we only provide an outline. First, note that the user selection is decoupled among different groups. Let M_d be the number of eligible transmitters for Group d, then M_d is binomially distributed with parameter $(\beta_d\, n,\, p')$ (similar to (5.27)). In this case, G_{sum} is a mixture of sums of order statistics described by Lemma 5.3.1, i.e., $G_{sum} = \sum_{d=1}^{D} S_{M_d}^{\beta_d k_s}(\rho_d)$ in distribution given M_d sufficiently large. The rest of the proof is similar to Theorem 5.3.2.

With slight modification of Lemma 5.3.3, we choose the number of active secondary transmitters as:

$$k_s^* = \underbrace{\sqrt{\frac{\Gamma}{eP\sum_{d=1}^{D} \beta_d \lambda_d}}}_{c'}\,(n)^{\frac{1}{2}}. \tag{5.19}$$

The above equation indicates that as the *average* cross-channel gain $\sum_d \beta_d \lambda_d$ increases, fewer secondary transmitters should be activated simultaneously. Notice that k_s^* becomes identical to that given by Lemma 5.3.3 (with $N_p = 1$) when $\lambda_d = 1$ for $1 \leq d \leq D$. Based this choice of k_s and Lemma 5.4.1, we obtain the following results.

Theorem 5.4.2 *For the network described above, the average secondary throughput R_{mac} satisfies*

$$R_{mac} \geq \frac{1}{2}\log n + C_2 + O(\frac{1}{\log n}), \tag{5.20}$$

$$R_{mac} \leq \frac{1}{2}\log n + C_2 + C_0 + O(\frac{1}{\log n}), \tag{5.21}$$

for sufficiently large n by activating $k_s = c'\sqrt{n}$ transmitters, where $C_2 = \log \frac{\sum_d \beta_d \rho_d}{\sqrt{\sum_d \beta_d \lambda_d}} + \log \frac{2}{1+P_p}\sqrt{\frac{\Gamma P}{e}}$.

Proof The theorem follows by substituting $k_s = c'\sqrt{n}$ into Lemma 5.4.1.

Remark 5.4.1 *From Theorem 5.4.2, the growth rate of R_{mac} is $\frac{1}{2}\log n$, which is optimal and thus is unaffected due to the imposition of the fairness constraint. Besides the growth rate, the impact of channel heterogeneity on the secondary throughput can also be seen by inspecting C_2: The lower (upper) bound of the throughput increases with the* average *secondary-channel gain, $\sum_d \beta_d \rho_d$, but decreases with the* average *cross-channel gain, $\sum_d \beta_d \lambda_d$. Intuitively, as $\sum_d \beta_d \lambda_d$ increases, statistically, the secondary transmitters more easily cause interference on the primary, thus fewer of them can be active simultaneously, which in turn leads to a smaller secondary throughput. Finally, note that Theorem 5.4.2 includes, as a special case, the results of Theorem 5.3.4 when the primary system simply consists of one transmitter-receiver pair.*

5.5 Numerical Results

In this section, we illustrate our results with simulations. We use $P_p = P = 10$ and $M_p = N_p = 1$. Unless otherwise specified, the allowable interference power on the primary receiver is $\Gamma = 5$. All simulations are averaged over 2×10^4 channel realizations.

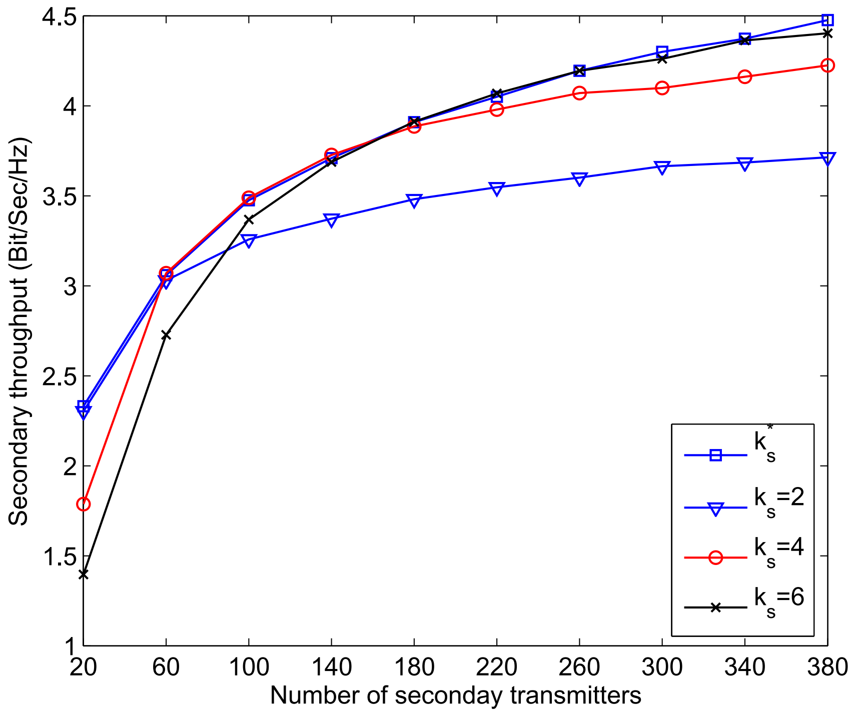

Figure 5.2. Optimal number of active secondary transmitters.

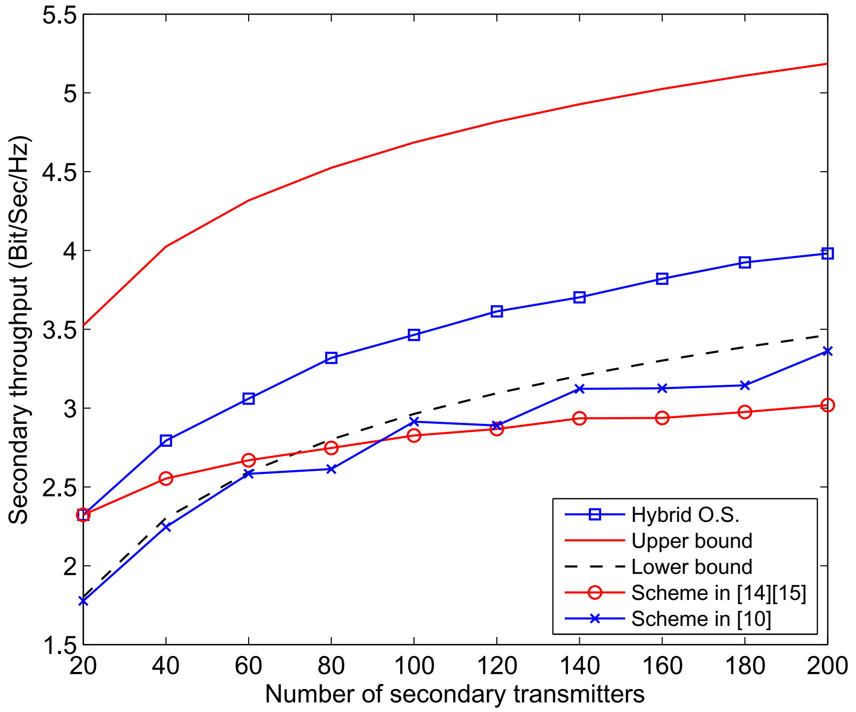

Figure 5.3. Throughput of Hybrid Opportunistic Scheduling and other schemes.

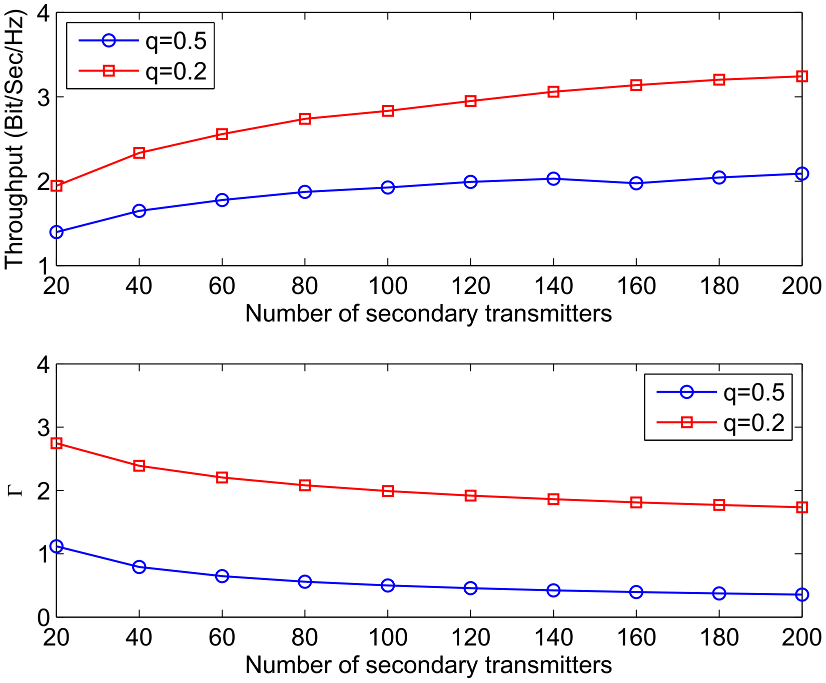

Figure 5.4. Throughput versus transmitter number, vanishing Γ

Figure 5.2 shows the (asymptotically) optimal number of active secondary transmitters characterized by Lemma 5.3.3. The throughput achieved by activating $k_s^* = \lceil\sqrt{\frac{\Gamma}{Pe}n}\rceil$ transmitters surpasses (or equals) that achieved by activating fixed k_s transmitters, for n from 20 to around 400. Although Lemma 5.3.3 only suggests k_s^{opt} cannot be far away from k_s^*, simulations imply that k_s^* may be indeed optimal. Intuitively, as n increases, the number of secondary transmitters that have desirably small cross-channel gains also increases on average, therefore, more secondary transmitters should be active simultaneously.

Figure 5.3 illustrates Theorem 5.3.4 and compares Hybrid Opportunistic Scheduling with several other schemes. The throughput of the proposed method is bounded by the asymptotic bounds in Theorem 5.3.4, even for small n. Hybrid Opportunistic Scheduling attains a throughput higher than that attained in Chapter 4, where the transmitters are selected only based on cross-channels without considering the secondary-channel conditions. Also, the achieved throughput is higher than that achieved in [56, 57] where the (single) secondary

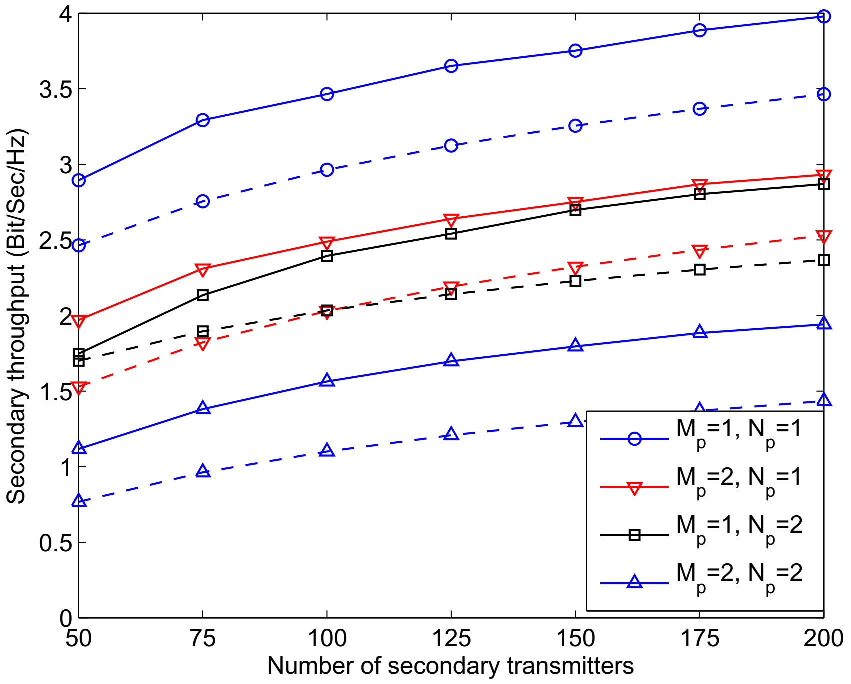

Figure 5.5. Throughput of Hybrid Opportunistic Scheduling under different M_p and N_p.

transmitter with the highest SINR is activated. The throughput of the proposed method scales as $\Theta(\log n)$, which is faster than the $\Theta(\log \log n)$ growth achieved in [56, 57].

Figure 5.5 shows the impact on the secondary throughput of the primary network (M_p and N_p). The dash lines correspond to the asymptotic lower bound derived by Theorem 5.3.4. As M_p increases, due to experiencing more interference from the primary, the secondary throughput decreases. As N_p increases, due to more constraints imposed by the primary, the secondary throughput again decreases.

The results of Corollary 5.3.5 are illustrated by Figure 5.4. The allowable interference power Γ declines (to zero) as n^{-q}, while the throughput still grows logarithmically with n. In addition, one can see the tradeoff given by Corollary 5.3.5: For larger q, the interference power decreases faster but the secondary throughput increases more slowly, and vice versa.

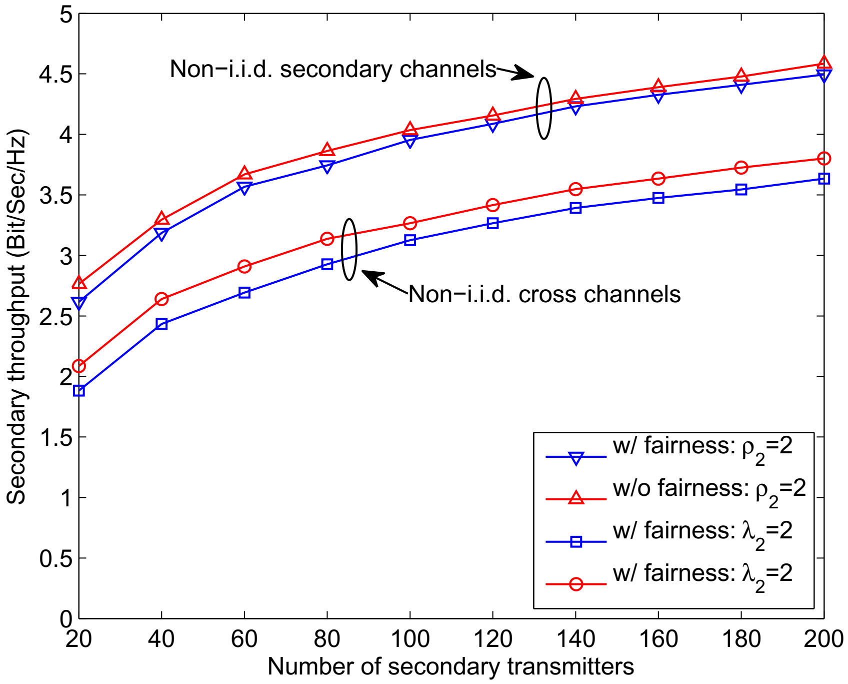

Figure 5.6. Throughput of Hybrid Opportunistic Scheduling under a fairness constraint.

Figure 5.6 and Figure 5.7 show the performance of Hybrid Opportunistic Scheduling under non-i.i.d. links. Here, $n = 50$ and $k_s = 4$; $D = 2$ and $\beta_1 = \beta_2 = 0.5$, i.e., two groups with equal number of transmitters. Figure 5.6 shows the secondary throughput for the case of $\rho_1 = \lambda_1 = \lambda_2 = 1$ and $\rho_2 = 2$ (non-i.i.d. secondary-channels), and for the case of $\rho_1 = \rho_2 = \lambda_1 = 1$ and $\lambda_2 = 2$ (non-i.i.d. cross channels). With the fairness constraint, the modified Hybrid Opportunistic Scheduling still attains a throughput that is very close to that attained without any fairness restriction.

Figure 5.7 shows the ratio of average portion of active time of Group 1 and Group 2. If this ratio equals 1, each transmitter has an equal portion of active time and the system is temporally fair [59]. The larger the ratio, the larger portion of active time of Group 2 relative to that of Group 1. One can see that user fairness is ensured under the modified scheduling.

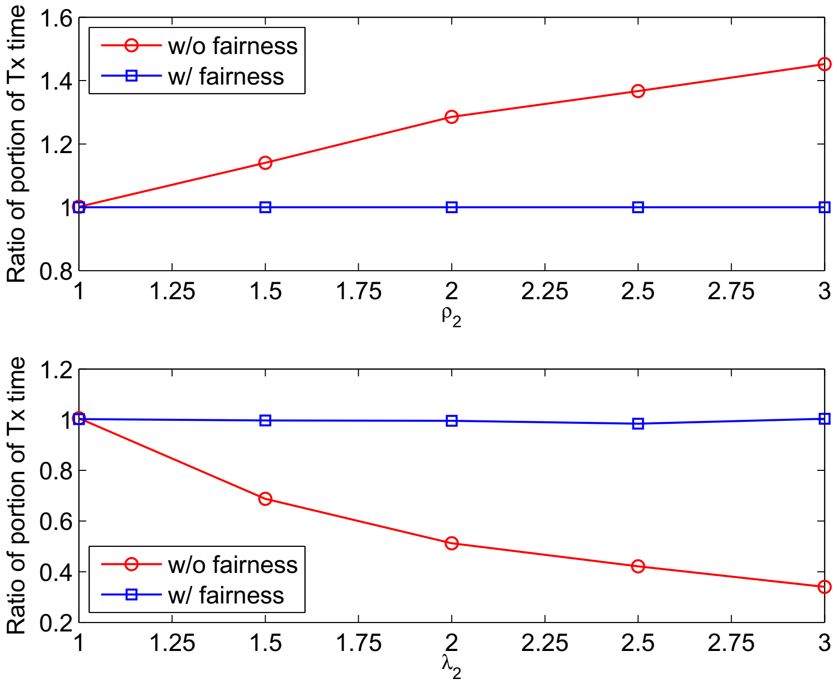

Figure 5.7. Fairness metric for Hybrid Opportunistic Scheduling.

5.6 Proof of Theorem and Lemma

5.6.1 Proof of Lemma 5.3.1

Proof Let $Z_1, \cdots, Z_b$ be i.i.d. exponentials with mean ρ. From [49], we know that $S_b^a(\rho)$ has the same distribution as

$$\sum_{i=1}^{b-a} \frac{a}{b-i+1} Z_i + Z_{b-a+1} + \cdots + Z_b. \tag{5.22}$$

Therefore we can calculate its expectation:

$$\mathbb{E}[S_b^a(\rho)] \triangleq \rho\mu_b^a = \rho a\bigg(\sum_{i=1}^{b} \frac{1}{i} - \sum_{i=1}^{a} \frac{1}{i}\bigg) + \rho a. \tag{5.23}$$

It is known [61] that, for any positive integer k,

$$\log k + \gamma + \frac{1}{2(k+1)} < \sum_{i=1}^{k} \frac{1}{i} < \log k + \gamma + \frac{1}{2k}, \tag{5.24}$$

where γ is the Euler constant. Hence, for sufficiently large b and a ($b \geq a$), we obtain

$$\mu_b^a = a \log \frac{b}{a} + a + O(1). \tag{5.25}$$

Now, we calculate the variance of $S_b^a(\rho)$. From (5.22), we have:

$$Var[S_b^a(\rho)] < (\rho a)^2 \sum_{i=a+1}^{b} \frac{1}{(i-1)i} + \rho^2 a < 2\rho^2 a. \tag{5.26}$$

Applying the Chebyshev inequality, for any $0 < \epsilon < 1$, we have

$$\mathbb{P}\bigg(\big|S_b^a(\rho) - \rho\mu_b^a\big| > \epsilon\rho\mu_b^a\bigg) < \frac{Var[S_b^a(\rho)]}{\big(\epsilon\rho\mu_b^a\big)^2} < O\bigg(\frac{1}{(\log b)^2}\bigg).$$

The above second inequality holds for any $a = O(b^\delta)$ and $\delta < 1$. The lemma follows by taking the complement of the random event in inequality.

5.6.2 Proof of Theorem 5.3.2

Proof To begin with, note that M (the size of $\mathcal{A}$) is binomially distributed with parameter

$$p = \left(1 - e^{-\frac{\alpha}{P}}\right)^{N_p}, \tag{5.27}$$

since $\{|g_{ji}|^2\}$ are i.i.d. exponentials with unit mean. For any $0 < \epsilon_1 < 1$, we have

$$\mathbb{P}\Big(|M - np| > \epsilon_1 np \Big) < \frac{(1-p)}{\epsilon_1^2 p\, n} = O(\frac{1}{n}) \tag{5.28}$$

based on the Chebyshev inequality. For convenience, we denote

$$n_1 = \lfloor (1-\epsilon_1)np \rfloor, \qquad n_2 = \lceil (1+\epsilon_1)np \rceil. \tag{5.29}$$

Then, from (5.28), we have

$$\mathbb{P}(M \geq n_1) > 1 - O(\frac{1}{n}), \quad \mathbb{P}(M \geq n_2) < O(\frac{1}{n}). \tag{5.30}$$

Now, we establish a lower bound. Based on (5.4), R_{mac} depends two independent random variables I_p and G_{sum}, where I_p is distributed as Gamma(M_p, P_p), and given $M = m$, G_{sum} has the same distribution as $S_m^{k_s}(P)$ for $m \geq k_s$ (see Lemma 5.3.1). Condition on $I_p = x$ and expand the conditional throughput $R_{mac|I_p}(x)$:

$$\begin{aligned} R_{mac|I_p}(x) =& \sum_{m=1}^{n} \mathbb{E}\left[\log\left(1 + \frac{G_{sum}}{1+x}\right) \middle| M = m \right] \mathbb{P}\big(M = m\big) \\ \geq& \sum_{m=n_1}^{n} \mathbb{E}\left[\log\left(1 + \frac{S_m^{k_s}(P)}{1+x}\right)\right] \mathbb{P}\big(M = m\big), \end{aligned} \tag{5.31}$$

where the inequality holds since we discard non-negative terms associated with $m < n_1$ in the summation and $n_1 > k_s$ for sufficiently large n. For any $0 < \epsilon < 1$, we further expand (5.31) by conditioning on the event $\mathcal{C}_m = \{S_m^{k_s}(P) > (1-\epsilon)P\mu_m^{k_s}\}$:

$$
\begin{aligned}
R_{mac|I_p}(x) &\geq \sum_{m=n_1}^{n} \mathbb{E}\left[\log\left(1+\frac{S_m^{k_s}(P)}{1+x}\right)\middle|\mathcal{C}_m\right]\mathbb{P}\big(M=m\big)\mathbb{P}\big(\mathcal{C}_m\big) \\
&> \sum_{m=n_1}^{n} \log\left(1+\frac{(1-\epsilon)P\mu_m^{k_s}}{1+x}\right)\mathbb{P}\big(M=m\big)\times\left(1-O\big(\frac{1}{(\log n)^2}\big)\right) && (5.32) \\
&> \log\left(1+\frac{(1-\epsilon)P\mu_{n_1}^{k_s}}{1+x}\right)\mathbb{P}\big(M\geq n_1\big)\left(1-O\big(\frac{1}{(\log n)^2}\big)\right) && (5.33) \\
&> \log\left(1+\frac{(1-\epsilon)P\mu_{n_1}^{k_s}}{1+x}\right)\big(1-O(\frac{1}{n})\big)\left(1-O\big(\frac{1}{(\log n)^2}\big)\right). && (5.34)
\end{aligned}
$$

To obtain (5.32), we use the result from Lemma 5.3.1 by noting $m = \Theta(n)$ for $m \geq n_1$:

$$
\mathbb{P}\big(\mathcal{C}_m\big) \geq 1 - O\big(\frac{1}{(\log n)^2}\big). \tag{5.35}
$$

We have (5.33), since $\mu_{n_1}^{k_s} \leq \mu_m^{k_s}$, $\forall\ m \geq n_1$. Finally, (5.34) uses (5.30).

From Lemma 5.3.1 and the fact that $n_1 = \Theta(n)$, we have $\mu_{n_1}^{k_s} = O(\log n)$. Since $\log(1+z) = \log z + \log(1+\frac{1}{z})$ for $z > 0$, we expand the right hand side of (5.34):

$$
R_{mac|I_p}(x) > \log\frac{P(1-\epsilon)\mu_{n_1}^{k_s}}{1+x} + O\big(\frac{1}{\log n}\big). \tag{5.36}
$$

Take expectation with respect to I_p and use the convexity of $h(z) = \log\big(1+\frac{c_1}{c_2+z}\big)$:

$$
\begin{aligned}
\mathbb{E}\big[R_{mac|I_p}(x)\big] &> \log\frac{P(1-\epsilon)\mu_{n_1}^{k_s}}{1+\mathbb{E}[I_p]} + O\big(\frac{1}{\log n}\big) \\
&= \log\frac{P\mu_{n_1}^{k_s}}{1+P_pM_p} + \log(1-\epsilon) + O\big(\frac{1}{\log n}\big). && (5.37)
\end{aligned}
$$

Finally, we calculate $\mu_{n_1}^{k_s}$. Since $\alpha = \frac{\Gamma}{k_s}$, from (5.27), we have $p \approx \big(\frac{\Gamma}{k_sP}\big)^{N_p}$ for large k_s. From Lemma 5.3.1, we have

$$
\mu_{n_1}^{k_s} = k_s\left(\log\frac{n\,\Gamma^{N_p}}{P^{N_p}\,k_s^{N_p+1}} + 1\right) + O(1). \tag{5.38}
$$

Substituting (5.38) into (5.37), and with some calculation, we have the desired lower bound in (5.10).

Now, we find an upper bound. Let $T = \frac{1}{1+I_p}$ and $R_{mac|T}(t)$ be the conditional throughput. Expand $R_{mac|T}(t)$ based on the event $\{M \le n_2\}$ and its complement:

$$\begin{aligned} R_{mac|T}(t) &= \mathbb{E}\Big[\log(1+tG_{sum}) \,\big|\, M \le n_2\Big]\mathbb{P}\big(M \le n_2\big) + \mathbb{E}\Big[\log(1+tG_{sum}) \,\big|\, M > n_2\Big]\mathbb{P}\big(M > n_2\big) \\ &\le \log\Big(1 + t\mathbb{E}\big[G_{sum} \,\big|\, M \le n_2\big]\Big) + \log\Big(1 + t\mathbb{E}\big[G_{sum} \,\big|\, M > n_2\big]\Big)\mathbb{P}\big(M > n_2\big), \qquad (5.39) \end{aligned}$$

where (5.39) uses the Jensen inequality. Since $\mathbb{E}[G_{sum}|M = i]$ is a non-decreasing function of i, we have

$$\begin{aligned} R_{mac|T}(t) &\le \log\big(1 + t\mathbb{E}[S_{n_2}^{k_s}(P)]\big) + \log\big(1 + t\mathbb{E}[S_n^{k_s}(P)]\big)\mathbb{P}\big(M > n_2\big) \\ &< \log\big(1 + tP\mu_{n_2}^{k_s}\big) + \log\big(1 + tP\mu_n^{k_s}\big)O\big(\frac{1}{n}\big), \qquad (5.40) \end{aligned}$$

where (5.40) uses (5.30). Take expectation with respect to T.

$$\begin{aligned} \mathbb{E}\big[R_{mac|T}(t)\big] &\le \log\big(1 + P\mu_{n_2}^{k_s}\,\mathbb{E}[T]\big) + \log\big(1 + P\mu_n^{k_s}\mathbb{E}[T]\big)O\big(\frac{1}{n}\big) \qquad (5.41) \\ &< \log\Big(1 + P\mu_{n_2}^{k_s}\,\mu_T\Big) + O\big(\frac{\log\log n}{n}\big) \qquad (5.42) \\ &= \log P\mu_{n_2}^{k_s} + \log\mu_T + \log\big(1 + O(1/\mu_{n_2}^{k_s})\big) + O\big(\frac{\log\log n}{n}\big), \qquad (5.43) \end{aligned}$$

where $\mu_T = \mathbb{E}[T]$. The Jensen inequality is used in (5.41) and the identity $\log(1+z) = \log z + \log(1+\frac{1}{z})$ for $z > 0$ is used in (5.43). Similar to (5.38), we have

$$\mu_{n_2}^{k_s} = k_s\Big(\log\frac{n\,\Gamma^{N_p}}{P^{N_p}\,k_s^{N_p+1}} + 1\Big) + O(1). \qquad (5.44)$$

Substituting (5.44) into (5.43), we obtain the desired upper bound in (5.11) with $C_0 = \log\big(\mu_T(1+M_pP_p)\big)$. Notice that $C_0 \ge 0$, because

$$\mu_T(1 + M_pP_p) = \mathbb{E}[1/(1+I_p)]\,\mathbb{E}[1+I_p] \ge 1, \qquad (5.45)$$

where the equality holds if and only if I_p is a constant.

5.6.3 Proof of Lemma 5.3.3

Proof The exact expression of R_{mac} as a function of k_s is unknown in Theorem 5.3.2, thus a direct maximization of $R_{mac}(k_s)$ is impossible. The idea of this proof is to (approximately) optimize bounds on R_{mac} and show that the resulting answer is sufficient for our purposes. We begin with the lower and upper bounds in Theorem 5.3.2, denoted as $L(k_s)$ and $U(k_s)$, which can be written as (ignoring vanishing terms):

$$L(k_s) = \log r(k_s) + \log \frac{P}{1 + M_p P_p}, \tag{5.46}$$

$$U(k_s) = \log r(k_s) + \log \frac{P}{1 + M_p P_p} + C_0, \tag{5.47}$$

where

$$r(k_s) = k_s \bigg(- (N_p + 1) \log k_s + \log n \big(\Gamma / P\big)^{N_p} + 1 \bigg). \tag{5.48}$$

Notice that $L(k_s)$ and $U(k_s)$ are identical function of k_s except a constant gap C_0. Intuitively, the value of k_s that maximizes $L(k_s)$ (or $U(k_s)$), denoted by k_s^*, should also (almost) maximize $R_{mac}(k_s)$. We justify this intuition in the rest of the proof.

First, we find k_s^*. Since $\log(\cdot)$ is a monotonic-increasing function, we maximize $r(k_s)$ instead. For the asymptotic analysis, k_s can be considered as a continuous variable. So, solving $r'(\cdot) = 0$, we obtain:

$$k_s^* = \big(\frac{\Gamma}{Pe}\big)^{\frac{N_p}{N_p+1}} n^{\frac{1}{N_p+1}}. \tag{5.49}$$

Now, consider $k_s = k_s^1$ such that $\log r(k_s^1) + C_0 < \log r(k_s^*)$. Then, k_s^1 is *not* the maximizer of $R_{mac}(k_s)$, because in this case $U(k_s^1) < L(k_s^*)$, which implies $R_{mac}(k_s^1) < R_{mac}(k_s^*)$. Therefore, k_s^{opt}, the true maximizer of $R_{mac}(k_s)$, must satisfy

$$\log r(k_s^{opt}) + C_0 \geq \log r(k_s^*). \tag{5.50}$$

Let $k_s^{opt} = k_s^* + \Delta k$. From (5.50), with some algebra, we have:

$$|\Delta k| \leq \sqrt{1 - \xi}\, k_s^*, \tag{5.51}$$

where

$$\xi = \exp(-C_0) = \big(\mu_T (1 + P_p M_p)\big)^{-1}. \tag{5.52}$$

Numerically, one can see that $\xi \approx 1$ and thus $\Delta k \approx 0$. For example, if $P_p = 10$, $\xi \approx 0.8$ for $M_p = 4$ and $\xi \approx 0.9$ for $M_p = 8$.

5.6.4 Proof of Theorem 5.3.6

Proof Consider an arbitrary $\mathcal{S}$ and $\{P_i\}_{i\in\mathcal{S}}$ that comply with the interference constraints imposed by the primary. We first enlarge the secondary throughput by assuming zero interference from the primary:

$$R_{mac} \leq \log\big(1 + \sum_{i\in\mathcal{S}} P_i|h_i|^2\big) \tag{5.53}$$

$$\leq \log\big(1 + G_{max}P_{sum}\big), \tag{5.54}$$

where

$$P_{sum} = \sum_{i\in\mathcal{S}} P_i, \qquad G_{max} = \max_{1\leq i\leq n} |h_i|^2. \tag{5.55}$$

Now we find an upper bound for R_{mac} regardless of transmission strategies. First, we bound P_{sum} and formulate an optimization problem:

$$\begin{aligned} &\max_{\mathcal{S},\{P_i\}} \ P_{sum} \\ s.t.: &\sum_{i\in\mathcal{S}} P_i|g_{ji}|^2 \leq \Gamma \text{ for } 1\leq j\leq N_p, \text{ and } P_i \leq P. \end{aligned} \tag{5.56}$$

which is a standard linear programming whose solution is denoted by P^*_{sum}. Here, P^*_{sum} is a random variable depending on the channel realizations. A direct solution requires joint optimization over $\mathcal{S}$ and $\{P_i\}$, but a simpler analysis exists for upper bounds. We relax the set of interference constraints in (5.56) to a single sum constraint, which never decreases P^*_{sum}:

$$\sum_{i\in\mathcal{S}} P_i I_{sum,i} \leq N_p\Gamma, \tag{5.57}$$

where

$$I_{sum,i} = \sum_{j=1}^{N_p} |g_{ji}|^2 \tag{5.58}$$

is the total cross-channel gains from the secondary transmitter i to all the primary receivers. Thus, $\{I_{sum,i}\}_{i=1}^n$ are i.i.d. Gamma$(N_p, 1)$. We order $I_{sum,i}$ among all the secondary transmitters:

$$\tilde{I}_{sum,1} \leq \cdots \leq \tilde{I}_{sum,n}. \tag{5.59}$$

Then, we construct the following problem by further relaxing the constraint of (5.57):

$$\begin{aligned} &\max_{\mathcal{S},\{P_i\}} P_{sum} \\ s.t.: &\sum_{i=1}^{|\mathcal{S}|} P_i \tilde{I}_{sum,i} \leq N_p \Gamma \text{ and } P_i \leq P. \end{aligned} \tag{5.60}$$

The solution for the above problem, denoted by $P^*_{sum,1}$, is always greater than or equal to P^*_{sum}. The corresponding $\{P_i\}$ achieves $P^*_{sum,1}$ are in form of $P_i \geq P_j$ for $i \leq j$. Thus, we have

$$P^*_{sum,1} \leq P N_{smax} \tag{5.61}$$

where N_{smax} is the maximum possible value of $|\mathcal{S}|$ that satisfies

$$P \sum_{i=1}^{|\mathcal{S}|-1} \tilde{I}_{sum,i} \leq N_p \Gamma. \tag{5.62}$$

For brevity, we outline the rest of the proof. It can be shown that N_{smax} converges to $\Theta(n^{\frac{1}{N_p+1}})$ in probability. Because $P^*_{sum} \leq PN_s$ and G_{max} (the maxima of n i.i.d. exponentials) scales as $\log n$ [39, 49], we have (see (5.54)):

$$\begin{aligned} R^{opt}_{mac} &\leq \log\left(\Theta(n^{\frac{1}{N_p+1}}) \log n\right) \qquad (5.63) \\ &= \frac{1}{N_p+1} \log n + O(\log \log n). \qquad (5.64) \end{aligned}$$

CHAPTER 6
SPECTRUM SHARING WITH DISTRIBUTED RELAY SELECTION AND CLUSTERING

6.1 Introduction

Spectrum-sharing [35, 38] allows unlicensed (secondary) users to share the spectrum of licensed (primary) users as long as the interference caused on the primary is tolerable. This problem is often formulated as maximizing the secondary rate subject to interference constraints on the primary, or as the dual problem of minimizing the interference on the primary subject to a fixed rate for the secondary. Thus, reducing the interference footprint of the secondary is of paramount interest in spectrum sharing. Multihop relaying and cooperative communication is known to significantly mitigate interference and increase the sum-throughput in many multi-user scenarios [2], among others in broadcast channels [3], multiple access channels [4] and interference channels [5]. This has motivated the use of relays in spectrum sharing networks [6–12].

This chapter studies a spectrum sharing network consisting of multiple primary nodes and a secondary system with M-antenna source and destination, and n half-duplex relays. Unlike conventional relay networks [45,62], the secondary relays must not only maximize the secondary rate but also control the interference on the primary, thus new cooperative algorithms are called for. To achieve this goal we propose and investigate an approach involving amplify-and-forward (AF) relaying as well as relay selection. Under the proposed framework a closed-form expression is derived for the secondary rate, showing that it increases as $(M \log n)/2$. Furthermore, we propose an augmented scheduling algorithm that recovers the half-duplex loss and improves the constant factor in the throughput growth rate. Finally, we characterize the trade-off between the secondary rate and the primary interference, showing that the interference on the primary can be reduced asymptotically to zero while the secondary rate still grows logarithmically with n. Our results suggest that to maximize the secondary rate subject to primary interference constraints, one must activate a subset of relays that are chosen based on their interference profile on the primary, each of the relays transmit with power inversely proportional to n, and the secondary source must operate at a power level potentially below its maximum available power. These outcomes are unique to the cognitive relay networks and are distinct from the conventional relay networks, e.g., [45].

Some of the related work is as follows. Zhang et al. [37] studied the secondary power allocation under various power and interference constraints. The throughput limits of spectrum-sharing broadcast and MAC were analyzed in [52] (Chapter 4). Recently, relaying in spectrum sharing networks has attracted attention. For secondary outage probability Zou et al. [11] and Lee et al. [12] proved that the relay selection in spectrum-sharing achieved the same diversity as conventional relay networks. For decode-and-forward (DF) relaying, Mietzner et al. [7] studied power allocation subject to a desired secondary rate, and Asghari and Aissa [8] analyzed symbol error rate with relay selection. For AF-relaying, Li et al. [9] selected a single relay to maximize the secondary rate, and Naeem et al. [10] numerically analyzed a greedy relay selection algorithm.

6.2 System Model

We consider a spectrum sharing network consists of N_p primary nodes and a secondary system with an M-antenna source, an M-antenna destination and n single-antenna half-duplex relays, as shown in Figure 6.1. The average interference power caused by the secondary on each of the primary nodes must be less than γ [63]. Let $\mathbf{H} \in \mathcal{C}^{M\times n}$ be the channel coefficient matrix from the source to the relays, and $\mathbf{F} \in \mathcal{C}^{n\times M}$ and $\mathbf{G} \in \mathcal{C}^{n\times N_p}$ be the channel coefficient matrices from the relays to the destination and the primary nodes, respectively. Denote $\mathbf{h}_{p,\ell} \in \mathcal{C}^{M\times 1}$ as the channel vector from the source to the primary node ℓ, $1 \leq \ell \leq N_p$. The source has no direct link to the destination, a widely used model [8, 45, 64] appropriate for geometries where the relays are roughly located in the middle of the source and destination.

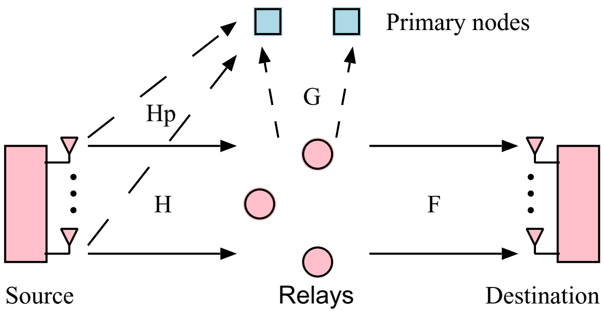

Figure 6.1. System model of Chapter 6

A block-fading model is considered where all entries of $\mathbf{H}$, $\mathbf{F}$, $\mathbf{G}$ and $\mathbf{h}_{p,\ell}$ are zero-mean i.i.d. circular symmetric complex Gaussian ($\mathcal{CN}$) with variance σ_s^2, σ_d^2, σ_p^2 and σ_{sp}^2, respectively.

The source communicates with the destination via two hops, which in general lowers the required transmit power and thus reduces the interference on the primary. In the first hop, the source sends M independent data streams across M antennas with equal power. The relay i receives

$$r_i = \sqrt{\frac{P_s}{M}}\mathbf{h}_i^t\mathbf{s} + n_i, \tag{6.1}$$

where P_s is the source transmit power, which must be less than a power constraint $\bar{P}_s$, $\mathbf{s} \in \mathcal{C}^{M\times 1}$ is i.i.d. Gaussian signals, $\mathbf{h}_i^t \in \mathcal{C}^{1\times M}$ is the row i of $\mathbf{H}$, namely the channel vector between the relay i and the source, and n_i is additive noise with distribution $\mathcal{CN}(0,1)$.

In the second hop, a subset of the relays is selected to transmit to the destination. We define a random variable T_i to indicate whether the relay i is selected (eligible):

$$T_i = \begin{cases} 1, & \text{the relay } i \text{ is eligible} \\ 0, & \text{otherwise} \end{cases}. \tag{6.2}$$

No cooperation among the relays is allowed due to their distributed nature. Each relay rotates and scales r_i by

$$c_i = e^{j\theta_i}\sqrt{\frac{P_r}{\mathbb{E}[T_i](P_s\sigma_s^2+1)}}. \tag{6.3}$$

where P_r is the average relay power and θ_i is the rotation angle, which are designed in the sequel. Therefore, the signal transmitted by the relay i is $T_i c_i\, r_i$.

After the relay forwarding, the received signal vector at the destination is

$$\mathbf{y} = \sqrt{\frac{P_s}{M}}\underbrace{\mathbf{FDH}}_{\tilde{\mathbf{H}}}\mathbf{s} + \underbrace{\mathbf{FDn} + \mathbf{w}}_{\tilde{\mathbf{w}}}, \tag{6.4}$$

where $\mathbf{D} = \mathrm{diag}(T_1 c_1, \cdots, T_n c_n)$ is the relay processing matrix and $\tilde{\mathbf{w}}$ is the equivalent additive noise. The equivalent channel matrix $\tilde{\mathbf{H}}$ has entries

$$[\tilde{\mathbf{H}}]_{mq} = \sum_{i=1}^{n} T_i \, c_i \, f_{mi} \, h_{iq}, \tag{6.5}$$

where f_{mi} and h_{iq} are $[\mathbf{F}]_{mi}$ and $[\mathbf{H}]_{iq}$, respectively.

In this chapter, we focus on the effect of the number of relays on the secondary rate, i.e., the so-called "scaling laws" for the relays in a spectrum-sharing system. Thus, we allow n to increase while N_p remains bounded. Analysis of scaling laws has a long and established history in wireless communications. Among the many examples we mention a few, e.g., [34, 45, 65].

We refer to cross channels between secondary transmitters and primary receivers as *interference links.* We assume the destination knows $\mathbf{F}$, $\mathbf{D}$ and $\mathbf{H}$, and the relays only know the instantaneous channel gains to which they directly connect, i.e., $\mathbf{h}_i$ and the column i of $\mathbf{F}$. The interference (thus the channels) from the primary to the secondary is not explicitly modeled for brevity, because its impact can be absorbed into the noise term $\tilde{\mathbf{w}}$.

The cross-channel CSI requirements in a TDD system can be met by the secondary nodes detecting packets emitted from the primary nodes. Otherwise, under the spectrum leasing model [58], the primary nodes can be expected to actively promote spectrum reuse by transmitting pilots that can be used for cross-channel gain estimation. The latter model applies to both TDD and FDD. Regarding the precision of cross-channel CSI, only the magnitude of the channel gains are needed, and the system can be made robust to imperfections in the cross-channel CSI to the relays, as shown in subsequent discussions (see Remark 6.3.1).

6.3 Spectrum-Sharing with relay selection and clustering

Relays that have weak interference links but strong secondary links are useful for spectrum sharing, while relays that produce a strong interference on the primary may do more harm than good. Therefore we use relay selection. In spectrum sharing, relay selection and allocation of transmit powers are coupled through the interference constraint, an issue that is not encountered in conventional (non-spectrum sharing) relaying. To make the problem tractable, we propose a two-step approach: first the allowable interference per relay is bounded, leading to the creation of an eligible relay set. Then the secondary rate is maximized by selecting appropriate relays from among the eligible set and coordinating their transmissions in a manner shown in the sequel.

6.3.1 Eligible Relay Selection

The interference on the primary nodes is controlled by activating only the relays with weak interference links. We design the relay selection in a distributed manner that does not require CSI exchange among the relays. A relay is eligible if and only if all of its own interference link gains are less than a pre-designed threshold α. So from (6.2)

$$T_i = \begin{cases} 1, & |g_{\ell i}|^2 \leq \alpha \text{ for } \ell = 1, \cdots, N_p \\ 0, & \text{otherwise} \end{cases}, \tag{6.6}$$

where $|g_{\ell i}|^2$ is the channel gain between the relay i and the primary node ℓ. Note that $\{|g_{\ell i}|^2\}_{\ell,i}$ are i.i.d. exponentials with mean σ_p^2, so $\{T_i\}_i$ are i.i.d. Bernoulli random variables with success probability

$$p = (1 - e^{-\alpha/\sigma_p^2})^{N_p}. \tag{6.7}$$

Since each relay determines eligibility based on its own interference links, the eligible relay selection is independent across the relays. The average interference from the secondary system to the primary node ℓ is

$$\gamma_\ell = \frac{1}{2}\mathbb{E}\big[(\sum_{i=1}^{n} g_{\ell i} t_i)(\sum_{i=1}^{n} g_{\ell i}^* t_i^*)\big] + \frac{P_s}{2M}\mathbb{E}\big[|\mathbf{h}_{p,\ell}|^2\big] \tag{6.8}$$

$$= \frac{P_r}{2}\sum_{i=1}^{n}\mathbb{E}\big[|g_{\ell i}|^2 \big| T_i = 1\big] + \frac{\sigma_{sp}^2 P_s}{2}, \tag{6.9}$$

where the factor $\frac{1}{2}$ is due to the fact that the relays and the source only transmit during half of the time. The second equality holds since the design of θ_i is independent of interference links, as shown soon. Since $T_i = 1$ implies $|g_{\ell i}|^2 \leq \alpha \ \forall \ell$, we have

$$\mathbb{E}\big[|g_{\ell i}|^2 \big| T_i = 1\big] < \int_0^\alpha \frac{xe^{-x/\sigma_p^2}}{\sigma_p^2}dx = \sigma_p^2 - e^{-\alpha/\sigma_p^2}(\alpha + \sigma_p^2) \tag{6.10}$$

$$\triangleq f(\alpha). \tag{6.11}$$

Combining (6.9) and (6.11), we have $\forall \ell$, $\gamma_\ell \leq \gamma$ if α and P_r satisfy

$$nP_r f(\alpha) \leq \max(\gamma_r, 0), \tag{6.12}$$

where $\gamma_r = 2\gamma - \sigma_{sp}^2 P_s$. As long as (6.12) holds, the interference on all the primary nodes is ensured to be less than γ, although the relays are selected distributedly. In our two-hop communication the source power P_s is chosen so that $\gamma_r > 0$, and otherwise the secondary rate is zero.

Remark 6.3.1 *We briefly discuss CSI uncertainty in the CSI of relay cross-channel gains. Denote the (relay) estimated cross channel gain as $|\hat{g}_{\ell i}|^2$. For simplicity, consider $|\hat{g}_{\ell i}|^2$ has the same exponential distribution as the true channel gain $|g_{\ell i}|^2$. Assume uncertainty can be modeled as an interval, e.g., that the true cross-channel gain is in the interval $[0, (1+\epsilon)|\hat{g}_{\ell i}|^2]$ for some known and fixed ϵ. In this case, if α and P_r satisfy*

$$nP_r f(\alpha + \epsilon) \leq \max(\gamma_r, 0),$$

the interference constraints on the primary will still be ensured. Since $f(\cdot)$ is an increasing and bounded function, the impact of uncertainty ϵ is to reduce the transmit power at the relays.

6.3.2 Distributed Relay Clustering

The second part of the proposed method aims to maximize the secondary rate. Recall that the source and destination have M antennas each; the relays are divided correspondingly into M groups $\{\mathcal{G}_m \ , \ 1 \leq m \leq M\}$, where each group of relays aims to provide a virtual pipe between one of the source antennas and the corresponding destination antenna. This channel-diagonalization approach is reminiscent of [45] but requires more sophisticated analysis because the (eligible) relay set is random, as shown in the sequel.

The relay $i \in \mathcal{G}_m$ rotates the received signal by θ_i such that

$$e^{j\theta_i} f_{mi} h_{im} = |f_{mi}||h_{im}|. \tag{6.13}$$

In this case, all the relays in $\mathcal{G}_m$ forward the signal sent by the source-antenna m coherently to the destination-antenna m.

Now, the challenge is to decide the assignment of relays to the group $\mathcal{G}_m$, for $1 \leq m \leq M$. We focus on distributed methods so that the coordination among relays is reduced. In addition, we decouple the relay clustering from the relay selection: the relays decide their groups according to their source-relay and relay-destination channels but independent of the interference links. Therefore, under this framework, $\{\theta_i\}_{i=1}^n$ and $\{T_i\}_{i=1}^n$ are mutually independent. This decoupling allows us to leverage existing relaying methods to enhance the secondary rate while bounding the primary interference. It also greatly simplifies the analysis.

We shall consider two clustering schemes:

Fixed Clustering

Here, each of the groups has n/M relays.[1] Subject to this condition, the relays are assigned to the groups in a pre-defined manner. Without loss of generality, we assume:

$$\mathcal{G}_m = \left\{ i : \frac{(m-1)n}{M} + 1 \leq i \leq \frac{mn}{M}, \ 1 \leq m \leq M. \right\}.$$

[1] We assume the number of relays n is so that n/M is an integer, however, this restriction is not essential and can be relaxed [45].

Gain Clustering

In this clustering we have

$$\mathcal{G}_m = \left\{ i : |h_{im}| > |h_{iq}|, \quad q \neq m,\, 1 \leq q \leq M \right\}.$$

In other words, the groups are assigned based on the relays' channel gain to source antennas. A relay (distributedly) decides to join in the group m if its gain to the m-th source antenna is the stronger than any other channel gains. The group assignment of relays is independent from each other and is also independent of relay eligibility. Note that $\mathcal{G}_m$ is no longer fixed but depends on the source-relay channels. Because all channels are i.i.d., a relay has equal probability of choosing any of the groups. Therefore $|\mathcal{G}_m|$ (the cardinality of $\mathcal{G}_m$) is binomially distributed with parameters $(n, \frac{1}{M})$.

6.4 Secondary Rate in Spectrum-sharing with Relays

We first derive a general closed-form expression for the secondary rate under the proposed framework, and then evaluate the achievable rate for specific methods.

6.4.1 Calculation of Secondary Rate

From (6.4), conditioned on $\mathbf{F}$, $\mathbf{D}$ and $\mathbf{H}$, $\tilde{\mathbf{w}}$ is a Gaussian vector with autocorrelation

$$\mathbf{W} = \mathbf{I} + \mathbf{F}\mathbf{D}\mathbf{D}^{\dagger}\mathbf{F}^{\dagger}. \tag{6.14}$$

The secondary rate in the presence of n relays is denoted with R_n and is given by:

$$R_n = \frac{1}{2} \log\det \left(\mathbf{I} + \frac{P_s}{M} \tilde{\mathbf{H}}\tilde{\mathbf{H}}^{\dagger}\mathbf{W}^{-1} \right), \tag{6.15}$$

where $\frac{1}{2}$ is due to the half-duplex relay constraint.

Now, we find R_n for large n. First, from (6.5) and (6.13), the entry of $\tilde{\mathbf{H}}$ is

$$[\tilde{\mathbf{H}}]_{mq} = \begin{cases} A_{mm} + B_{mm}, & q = m \\ C_{mq}, & q \neq m \end{cases}, \tag{6.16}$$

where

$$A_{mm} = \sqrt{\frac{P_r}{p(\sigma_s^2 P_s + 1)}} \sum_{i \in \mathcal{G}_m} T_i \, |f_{mi}| \, |h_{im}|,$$

$$B_{mm} = \sqrt{\frac{P_r}{p(\sigma_s^2 P_s + 1)}} \sum_{i \notin \mathcal{G}_m} T_i \, f_{mi} \, h_{im} \, e^{j\theta_i},$$

$$C_{mq} = \sqrt{\frac{P_r}{p(\sigma_s^2 P_s + 1)}} \sum_{i=1}^{n} T_i \, f_{mi} \, h_{iq} \, e^{j\theta_i}. \tag{6.17}$$

The terms in A_{mm}, B_{mm} and C_{mq} are mutually independent, because $\{T_i\}_{i=1}^n$ and $\{\theta_i\}_{i=1}^n$ are independent from each other. So we have the following lemma.

Lemma 6.4.1 *If* $\min_{1 \le m \le M} |\mathcal{G}_m| \xrightarrow{w.p.1} \infty$ *as* $n \to \infty$*, we have*

$$\frac{A_{mm}}{n} - \frac{1}{n}\sqrt{\frac{pP_r}{\sigma_s^2 P_s + 1}} \sum_{i \in \mathcal{G}_m} \mathbb{E}[|f_{mi}||h_{im}|] \xrightarrow{w.p.1} 0, \tag{6.18}$$

$$\frac{B_{mm}}{n} - \frac{1}{n}\sqrt{\frac{pP_r}{\sigma_s^2 P_s + 1}} \sum_{i \notin \mathcal{G}_m} \mathbb{E}[f_{mi} h_{im} e^{-j\theta_i}] \xrightarrow{w.p.1} 0, \tag{6.19}$$

$$\frac{C_{mq}}{n} - \frac{1}{n}\sqrt{\frac{pP_r}{\sigma_s^2 P_s + 1}} \sum_{i=1}^{n} \mathbb{E}[f_{mi} h_{iq} e^{-j\theta_i}] \xrightarrow{w.p.1} 0. \tag{6.20}$$

Proof The proof follows from [66, Theorem 2.1] and [43, Theorem 1.8.D], and is omitted here.

From Lemma 6.4.1, given $|\mathcal{G}_m| \xrightarrow{w.p.1} \infty \; \forall m$, we have:

$$\frac{\tilde{\mathbf{H}}}{n} - \mathrm{diag}(a_1, \cdots, a_M) \xrightarrow{w.p.1} 0, \tag{6.21}$$

where

$$a_m = \frac{1}{n}\sqrt{\frac{pP_r}{(\sigma_s^2 P_s + 1)}} \sum_{i \in \mathcal{G}_m} \mathbb{E}[|f_{mi}||h_{im}|]. \tag{6.22}$$

The above analysis indicates that $\tilde{\mathbf{H}}$ converges to a diagonal matrix for large n (with probability 1). We now show that $\mathbf{W}$ is also diagonalized as n increases. From (6.14), we have

$$[\mathbf{W}]_{mq} = \sum_{i=1}^{n} \frac{T_i P_r f_{mi} f_{iq}^*}{p(P_s \sigma_s^2 + 1)} + \delta_{mq}, \tag{6.23}$$

where $\delta_{mq} = 1$ if $m = q$ and $\delta_{mq} = 0$ if $m \neq q$. One can verify Kolmogorov conditions [43, Theorem 1.8.D], and therefore obtain

$$\frac{\left[\mathbf{W}\right]_{mq}}{n} - \frac{1}{n}\left(\sum_{i=1}^{n} \frac{P_r \mathbb{E}[f_{mi} f_{iq}^*]}{P_s \sigma_s^2 + 1} + \delta_{mq}\right) \overset{w.p.1}{\longrightarrow} 0, \tag{6.24}$$

where

$$\mathbb{E}[f_{mi} f_{iq}^*] = \begin{cases} \mathbb{E}[|f_{mi}|^2], & m = q \\ 0, & m \neq q \end{cases}. \tag{6.25}$$

Therefore, we have

$$\frac{\mathbf{W}}{n} - \mathrm{diag}(b_1, \cdots, b_M) \overset{w.p.1}{\longrightarrow} 0, \tag{6.26}$$

where

$$b_m = \frac{P_r \sum_{i=1}^{n} \mathbb{E}[|f_{mi}|^2]}{n(P_s \sigma_s^2 + 1)} + \frac{1}{n}. \tag{6.27}$$

From (6.21) and (6.26), for large n, the end-to-end channel between the source and the destination is approximately decoupled into M parallel channels under the proposed framework, where the channel coefficient m is a_m and the received noise has variance b_m. The capacity of this parallel channel is

$$\overline{R} = \frac{1}{2} \sum_{m=1}^{M} \log\left(1 + \frac{n P_s a_m^2}{M b_m}\right), \tag{6.28}$$

Therefore, it is reasonable to expect that $R_n \approx \overline{R}$ for large n. After some calculation (omitted for brevity), we obtain the following result.

Theorem 6.4.2 *Consider a secondary system with an M-antenna source, an M-antenna destination, and n single-antenna relays, in the presence of N primary nodes each tolerating interference no more than γ. The secondary rate satisfies*

$$R_n - \overline{R} \overset{w.p.1}{\longrightarrow} 0, \quad n \to \infty, \tag{6.29}$$

under the proposed relay selection and clustering framework.

6.4.2 Achievable Rate under Specific Clustering Schemes

We apply Theorem 6.4.2 to fixed clustering and gain clustering.

Fixed Clustering

In this scheme, $|\mathcal{G}_m| = \frac{n}{M}$ (so Lemma 6.4.1 is applicable), and $|f_{mi}|$ and $|h_{im}|$ are i.i.d. Rayleigh random variables with mean $\frac{\sigma_d\sqrt{\pi}}{2}$ and $\frac{\sigma_s\sqrt{\pi}}{2}$, respectively. Therefore, from (6.22), $a_m = \frac{\pi\sigma_s\sigma_d}{4M}\sqrt{\frac{pP_r}{\sigma_s^2 P_s+1}}$, for $1 \leq m \leq M$. Under this clustering, $|f_{mi}|^2$ is i.i.d. exponential with mean σ_d^2, and we have $b_m = \frac{\sigma_d^2 P_r}{\sigma_s^2 P_s+1} + \frac{1}{n}$, for $1 \leq m \leq M$. Substituting a_m and b_m into (6.28), $\overline{R}$ becomes

$$R^{(f)} = \frac{M}{2}\log\left(1 + \frac{np\pi^2\sigma_s^2\sigma_d^2 P_r P_s}{16M^3(\sigma_d^2 P_r + n^{-1}(\sigma_s^2 P_s + 1))}\right). \tag{6.30}$$

From Theorem 6.4.2, under fixed clustering, we have: $R_n - R^{(f)} \xrightarrow{w.p.1} 0$.

Gain Clustering

Since $|\mathcal{G}_m|$ is binomially distributed with parameters $(n, \frac{1}{M})$, we have $|\mathcal{G}_m|/n \xrightarrow{w.p.1} 1/M$, and Lemma 6.4.1 is again applicable. Due to the independence of $|f_{mi}|$ and $|h_{im}|$, from (6.22), we have

$$a_m = \frac{1}{n}\sqrt{\frac{pP_r}{(\sigma_s^2 P_s + 1)}}\sum_{i\in\mathcal{G}_m}\mathbb{E}[|f_{mi}|]\mathbb{E}[|h_{im}|]. \tag{6.31}$$

where $\mathbb{E}[|f_{mi}|] = \frac{\sigma_d\sqrt{\pi}}{2}$ (i.i.d. Rayleigh) and $\mathbb{E}[|h_{im}|] = \max_{1\leq m\leq M}|h_{mi}|$, which is the maximum of M i.i.d. Rayleigh random variables. We have

$$\begin{aligned}\mu_h &= \mathbb{E}\big[\max_{1\leq m\leq M}|h_{im}|\big] \\ &= \int_0^\infty \frac{2Mx^2}{\sigma_s^2}e^{-x^2/\sigma_s^2}\big(1 - e^{-x^2/\sigma_s^2}\big)^{M-1}dx \qquad (6.32)\\ &= \sum_{m=0}^{M-1}(-1)^{M-m-1}\tbinom{M-1}{m}\frac{\sigma_s M\Gamma(\frac{3}{2})}{(M-m)^{3/2}}. \qquad (6.33)\end{aligned}$$

Note that $\mu_h = \frac{\sigma_s\sqrt{\pi}}{2}$ for $M = 1$ (no selection is needed), which is identical to the fixed clustering. Based on (6.31) and $|\mathcal{G}_m|/n \overset{w.p.1}{\to} 1/M$, we have $a_m - \frac{\sigma_d\mu_h}{2M}\sqrt{\frac{p\pi P_r}{(\sigma_s^2 P_s+1)}} \xrightarrow{w.p.1} 0$.

Under this clustering, b_m remains the same as the fixed clustering case, since $|f_{mi}|$ is still i.i.d. Rayleigh for $i \in \mathcal{G}_m$, $\forall m$. Substituting a_m and b_m into (6.28), we have

$$R^{(g)} = \frac{M}{2} \log\left(1 + \frac{np\pi\mu_h^2\sigma_d^2 P_r P_s}{4M^3(\sigma_d^2 P_r + n^{-1}(\sigma_s^2 P_s + 1))}\right), \tag{6.34}$$

then: $R_n - R^{(g)} \xrightarrow{w.p.1} 0$.

6.5 Optimal Power Strategy for Spectrum-sharing with relays

In general, one may envision two competing philosophies for relay selection: (1) Allow only relays that have extremely weak interference links to the primary. Only very few relays will qualify but each of them can transmit at high power. (2) Allow a large number of relays to be activated. In this case the relay powers must be lowered because not all interference links are as "good" as the previous case.

The key question is: which approach is better? Should we use a few select relays with excellent interference profiles, or more relays operating at lower power? In this section, we optimize the threshold α, the relay power P_r and the source power P_s, while bounding the primary interference. The results of this section show that in general the balance tips in favor of having more eligible relays operating at low power.

6.5.1 Optimal Design of α and P_r

Consider a fixed P_s. Since α and P_r depend on each other via (6.12), given α the maximum P_r is

$$P_r = \frac{\gamma_r}{nf(\alpha)}. \tag{6.35}$$

Substituting (6.35) and (6.7) into (6.30) and (6.34) shows that $R^{(f)}$ and $R^{(g)}$ attain their maxima (as a function of α) at $\alpha = \alpha_0$ where:

$$\alpha_o = \arg\max_{\alpha} \frac{\gamma_r P_s(1 - e^{-\alpha/\sigma_p^2})^{N_p}}{\gamma_r \sigma_d^2 + (\sigma_s^2 P_s + 1)f(\alpha)}. \tag{6.36}$$

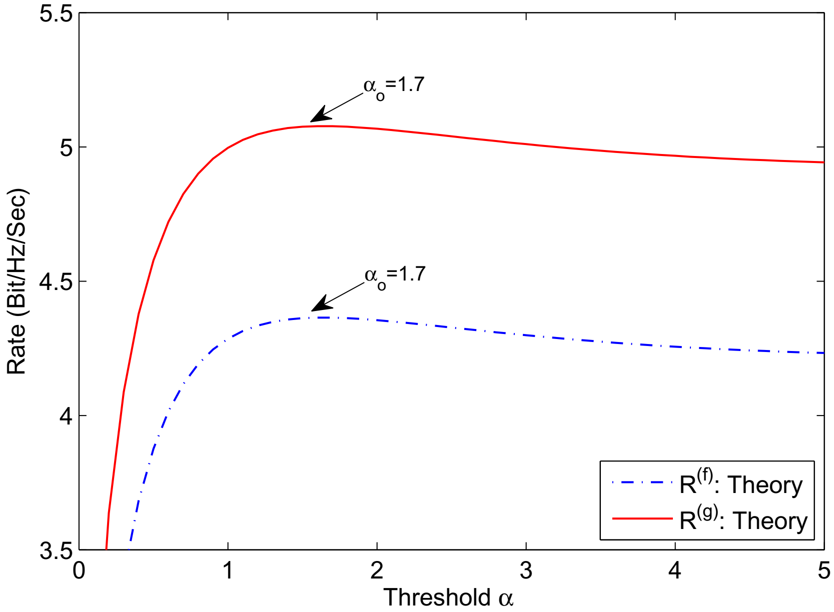

Figure 6.2. Optimal value of selection threshold α under $P_s = 5, n = 100$

A closed-form solution for α_o is unavailable but numerical solution can be easily obtained. Figure 6.2 shows the optimal design of α based on (6.36). For both fixed clustering and gain clustering, according to (6.36), $\alpha_o = 1.7$ maximizes the secondary rate.

Now, we characterize the asymptotic behavior of α_o, equivalently the optimal P_r. Because (6.36) is independent of n, the optimal threshold α is not a function of n. So from (6.35) the optimal average transmit power[2] is $P_r = \Theta(n^{-1})$, i.e., there exist real constants $d_1, d_2 > 0$ so that $d_1 n^{-1} \leq P_r \leq d_2 n^{-1}$. This implies that the secondary system should on average allow many relays to operate at low power. One may intuitively interpret this result as follows. To comply with the primary interference constraints, the sum power of relays must be bounded, and by spreading the total power among more relays better beamforming gain is achieved via coherent transmission.

Now, we study the scaling of the secondary rate. Consider examples with fixed clustering (Eq. (6.30)) and gain clustering (Eq. (6.34)). If $P_r = \Theta(n^{-1})$, for fixed α (not necessarily

[2] It can be shown that Theorem 6.4.2 still holds if P_r scales as $\Theta(n^{-1})$.

optimal), we have

$$R^{(f)} = \frac{M}{2}\log n + C_1, \quad R^{(g)} = \frac{M}{2}\log n + C_1 + C_2, \tag{6.37}$$

where

$$C_1 = \frac{M}{2}\log\frac{p\pi^2\sigma_s^2\sigma_d^2\gamma_r P_s}{16M^3(\sigma_s^2 P_s + 1)(\sigma_p^2\gamma_r + f(\alpha))}$$

and $C_2 = \log\frac{4\mu_h^2}{\pi\sigma_s^2}$. One can view C_2 as multi-antenna diversity gain by selecting over source-relay channels. From (6.37), the secondary rate increases as $(M\log n)/2$, which is summarized in the next theorem.

Theorem 6.5.1 *Consider a secondary system with an M-antenna source, an M-antenna destination, and n single-antenna relays, in the presence of N primary nodes each tolerating interference no more than γ. For $P_r = \Theta(n^{-1})$ and fixed α, the secondary rate satisfies*

$$\frac{R_n}{\frac{M}{2}\log n} \xrightarrow{w.p.1} 1, \tag{6.38}$$

under the proposed framework with both fixed clustering and gain clustering.

Theorem 6.5.1 holds for a broad class of clustering schemes, as long as the corresponding a_m and b_m are bounded but non-zero, i.e., the secondary end-to-end equivalent channel is diagonalized with probability 1 as n grows.

Remark 6.5.1 *It is possible to extend our results to the case of peak interference constraint γ. The secondary source will manage its instantaneous interference to be smaller than γ_s on all primary nodes by adjusting its transmit power according to the largest cross-channel gain to the primaries. Then, the sum interference from all the relays must be smaller than $\gamma_r = \gamma - \gamma_s$. Let $P_r = \xi/n$ where ξ is a positive constant. The instantaneous interference from all the relays to the primary node ℓ is $\gamma_\ell = \xi\sum_{i=1}^{n} T_i|g_{\ell i}|^2/n$. This implies that*

$$\gamma_\ell - \xi\,\mathbb{E}[T_i\,|g_{\ell i}|^2] \xrightarrow{w.p.1} 1$$

for an arbitrary $i \in \{1,\ldots,n\}$, where we have used the fact that T_i and $g_{\ell i}$ for all i have identical distributions. Therefore, $\xi = \gamma_r\big(\mathbb{E}[T_i\,|g_{\ell i}|^2]\big)^{-1}$ ensures the instantaneous interference on all the primary nodes to be smaller than γ with probability 1.

6.5.2 Optimal Source Power

Due to the primary interference constraints, for any chosen α the higher the source power P_s, the lower the relay power P_r, and vice versa. From (6.30) and (6.34), the rate-maximizing P_s is

$$P_s^* = \arg \max_{0<P_s<\min(\bar{P}_s, 2\gamma)} \frac{(2\gamma - \sigma_{sp}^2 P_s)P_s}{(2\gamma - \sigma_{sp}^2 P_s)\sigma_d^2 + (\sigma_s^2 P_s + 1)f(\alpha)}. \tag{6.39}$$

The unique solution of the above optimization problem is

$$P_s^* = \min(P_o, \bar{P}_s), \tag{6.40}$$

where $P_0 = \frac{\gamma}{\sigma_{sp}^2}$ if $\sigma_{sp}^2 \sigma_d^2 = \sigma_s^2 f(\alpha)$, otherwise:

$$P_o = \frac{2\gamma\sigma_d^2}{\sigma_{sp}^2\sigma_d^2 - \sigma_s^2 f(\alpha)} - \frac{\sqrt{\big(2\gamma\sigma_d^2 f(\alpha) + f^2(\alpha)\big)\big(\sigma_{sp}^2 + 2\gamma\sigma_s^2\big)}}{\sigma_{sp}\big(\sigma_{sp}^2\sigma_d^2 - \sigma_s^2 f(\alpha)\big)} \tag{6.41}$$

Figure 6.3 demonstrates the optimal source power as a function of three channel parameters σ_{sp}^2, σ_s^2 and σ_d^2. Three curves are shown, in each case one parameter varies while the other two are held constant (at unity). In this Figure $\bar{P}_s = 10$, $\gamma = 5$ and $f(\alpha) = 0.8$. As the source-primary channels become stronger, the source needs to reduce power; otherwise, the relay power must decrease to comply with the primary interference constraints, which curbs the rate achieved by the second hop. If the source-relay channels become stronger, the relay-destination links is the bottleneck and the relays need to transmit at higher power, thus once again the source needs to reduce power. In contrast, when the relay-destination channels become better, the source-relay channels are the bottleneck so the source needs to increase power.

6.5.3 Asymptotic Reduction of Interference on Primary

Multiple relays produce opportunities not only to enhance the secondary rate but also to reduce the interference on the primary. Suppose the interference on the primary nodes to be bounded as $\gamma = O\big(n^{-\delta}\big)$, which goes to zero as $n \to \infty$. From (6.12), it is sufficient to comply with this constraint if P_r decreases as $\Theta(n^{-(1+\delta)})$ and P_s decreases as $\Theta(n^{-\delta})$.

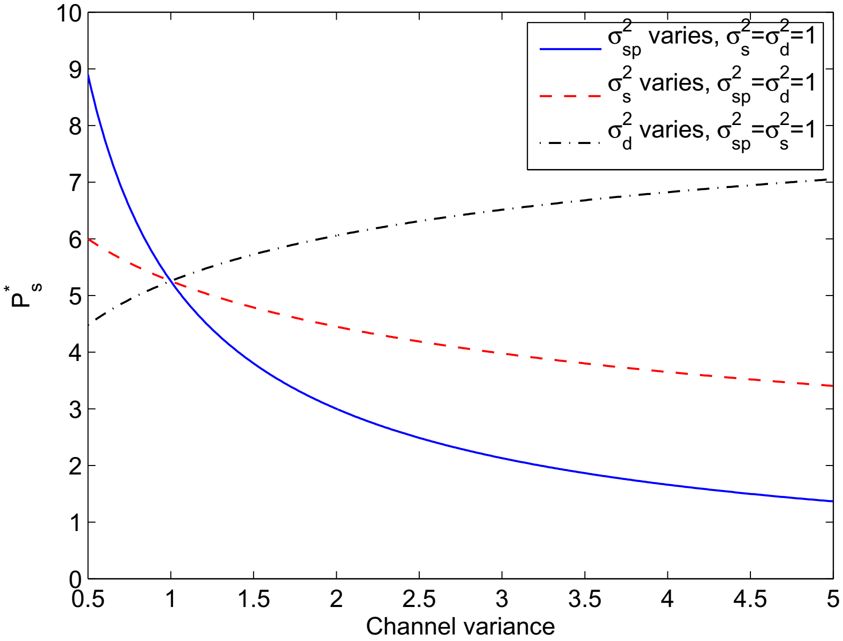

Figure 6.3. Optimal source power with $\gamma = 5, f(\alpha) = 0.8$

Substituting P_r and P_s into the expression of $R^{(f)}$ given by (6.30) and following some order calculation (the analysis of $R^{(g)}$ is the same thus omitted), we have

$$R^{(f)} = \begin{cases} \frac{M(1-2\delta)}{2} \log n + O(1) & \delta < \frac{1}{2} \\ O(1) & \delta \geq \frac{1}{2} \end{cases} \tag{6.42}$$

The above equation characterizes the trade-off between the secondary rate and the interference on the primary: the faster of the interference reduction, the slower of the rate growth. It also shows that the interference on the primary nodes may be mitigated (to zero asymptotically), while the secondary rate maintains to increase as $\Theta(\log n)$.

Remark 6.5.2 *In the above, the allowable interference γ is made to decline as $\Theta(n^{-\delta})$, which leads the growth rate to decrease linearly in δ. If γ is reduced more slowly, e.g., decreasing as $\Theta(\frac{1}{\log n})$, the secondary rate can increase at a rate of $\frac{M}{2}\log n$. If we try to mitigate the primary interference faster than $\Theta(1/\sqrt{n})$, the secondary rate will not increase logarithmically with n.*

6.6 Spectrum-sharing with Alternating Relay Protocol

In this section we consider issues raised by the relay half-duplex constraint, i.e., limitations that arise because relays cannot listen to the source at the same time as they are transmitting. When a subset of relays are activated for relaying the previously received information, the inactive relays are able to listen and receive information from the source, thus in principle the source can transmit continually and the half-duplex loss can be mitigated. This is the basic idea of *spectrum sharing with Alternating Relay Protocol*, which is the subject of this section.

The protocol consists of L transmission frames, as shown in Figure 6.4. It is assumed the channel coefficient remains constant during each frame, but varies independently from frame to frame. The source transmits during frames 1 through $L-1$, and remains silent during frame L. Since the source transmits $L-1$ data segments during L time intervals, the rate loss induced by the half-duplex relaying is a factor of $\frac{L-1}{L}$. The relays are partitioned into two groups $\mathcal{G}_1 = \{1 \leq i \leq \frac{n}{2}\}$ and $\mathcal{G}_2 = \{\frac{n}{2}+1 \leq i \leq n\}$. During even-numbered transmission frames a subset of the relays in $\mathcal{G}_1$ transmit to the destination, while the relays in $\mathcal{G}_2$ listen to the source. During odd-numbered transmission frames, a subset of the relays in $\mathcal{G}_2$ transmit, while the relays in $\mathcal{G}_1$ listen. As shown later, each of the two relay groups asymptotically achieves a rate that grows as $\frac{M}{2}\frac{L-1}{L}\log n$, thus the overall system has a rate that grows proportionally to $M\frac{L-1}{L}\log n$. Therefore a good part of the half-duplex rate loss can be recovered.

When either group $\mathcal{G}_1$ or group $\mathcal{G}_2$ is in the transmit mode, a subset of relays in the corresponding group is selected to transmit. A relay is selected (eligible) if its interference links satisfy (6.6), similar to Section 6.3.1. The average interference power on the primary node ℓ takes slightly different forms depending on whether L is even or odd. When L is even:

$$\gamma_\ell = \frac{(L-1)\sigma_{sp}^2 P_s}{L} + \frac{P_r}{2}\sum_{i\in\mathcal{G}_1}\mathbb{E}\big[|g_{\ell i}|^2\big|T_i=1\big] + \frac{(L-2)P_r}{2L}\sum_{i\in\mathcal{G}_2}\mathbb{E}\big[|g_{\ell i}|^2\big|T_i=1\big], \tag{6.43}$$

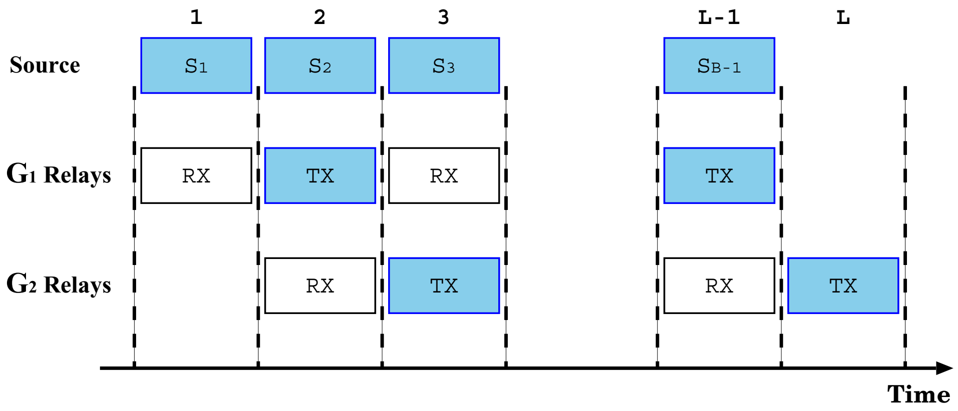

Figure 6.4. Transmission schedule in the alternating relay protocol (ARP)

and when L is odd:

$$\gamma_\ell = \frac{(L-1)\sigma_{sp}^2 P_s}{L} + \frac{(L-1)P_r}{2L}\left(\sum_{i\in\mathcal{G}_1}\mathbb{E}\left[|g_{\ell i}|^2\middle|T_i=1\right] + \sum_{i\in\mathcal{G}_2}\mathbb{E}\left[|g_{\ell i}|^2\middle|T_i=1\right]\right). \qquad (6.44)$$

To comply with the interference constraints on the primary nodes, the threshold α and the relay power P_r shall satisfy

$$nP_r f(\alpha) \leq \max(\gamma_L, 0), \qquad (6.45)$$

where $\gamma_L = \frac{2L}{L-1}\gamma - 2\sigma_{sp}^2 P_s$ with P_s so that $\gamma_L > 0$, and we use the fact that $\mathbb{E}\left[|g_{\ell i}|^2\middle|T_i = 1\right] = f(\alpha)$. Since from Section 6.5 the optimal P_r is proportional to n^{-1}, we let $P_r = \eta/n$, and re-write (6.45) as

$$\eta f(\alpha) \leq \max(\gamma_L, 0). \qquad (6.46)$$

For the Alternating Relay Protocol, relay clustering is accomplished in a manner similar to Section 6.3.2, therefore the details are omitted. During frame $2k$ (or $2k+1$), let $\mathcal{G}_{m,k}^{(1)} \subset \mathcal{G}_1$ (or $\mathcal{G}_{m,k}^{(2)} \subset \mathcal{G}_2$) be the set of relays that assists the antenna pair m. As long as $\min_{m,k,d} |\mathcal{G}_{m,k}^{(d)}| \to \infty$, the secondary rate will be obtained following the analysis similar to Section 6.4.

Remark 6.6.1 *At any point in time, it is possible to allow* all *non-transmitting relays to listen to the source, and be eligible to transmit in the next frame. This may give some gains, however, it also complicates the relay selection by introducing dependence between not only* *interference links but also other links such as source-relay and relay-relay links. It may be better for a relay even with a small interference on primary to remain inactive if it has also a weak channel to destination (therefore it cannot help much) but has a strong channel to the source (therefore it can listen well for the next round). Thus, any gains will come with a loss of elegance and tractability, and therefore this approach is not considered in this chapter.*

6.6.1 A Simple Example: $L = 3$

For illustration purposes, we consider $L = 3$, where $\mathcal{G}_1$ ($\mathcal{G}_2$) listen to the source during frame 1, and then transmit to the destination during frame 2 (frame 3). We assume fixed clustering is used with $|\mathcal{G}_{m,1}^{(d)}| = n/(2M)$, for $1 \leq m \leq M$ and $1 \leq d \leq 2$. Let $\mathbf{H}_d$ ($\mathbf{F}_d$) be the channel coefficient matrix between the relays in $\mathcal{G}_d$ and the source (the destination), and $\mathbf{H}_r$ be the channel coefficient matrix between $\mathcal{G}_1$ and $\mathcal{G}_2$ with i.i.d. $\mathcal{CN}(0, \sigma_r^2)$ entries.

We now analyze the rate achieved under Alternating Relay Protocol. The optimization of the threshold and the source power follows in a manner similar to Section 6.5 and thus is omitted here.

Rate Achieved by $\mathcal{G}_1$

After listening to the source at frame 1, $\mathcal{G}_1$ relays to the destination at frame 2. At the end of frame 2, similar to (6.4) the received signal at the destination is

$$\mathbf{y}_1 = \sqrt{\frac{P_s}{M}} \underbrace{\mathbf{F}_1 \mathbf{D}_1 \mathbf{H}_1}_{\tilde{\mathbf{H}}_1} \mathbf{s}_1 + \underbrace{\mathbf{F}_1 \mathbf{D}_1 \mathbf{n}_1 + \mathbf{w}_1}_{\tilde{\mathbf{w}}_1}, \tag{6.47}$$

where $\mathbf{s}_1$ is the signal sent by the source during frame 1, $\mathbf{n}_1$ is the noise forwarded by the group $\mathcal{G}_1$ of relays, $\mathbf{w}_1$ is the destination noise. For the group $\mathcal{G}_1$ the relay gains are collected into the relay processing matrix

$$\mathbf{D}_1 = \mathrm{diag}(T_1 c_1, \cdots, T_{\frac{n}{2}} c_{\frac{n}{2}}), \tag{6.48}$$

where c_i is given by (6.3) so that the average relay power constraints are satisfied. One can verify that the equivalent channel $\tilde{\mathbf{H}}_1$

$$\frac{\tilde{\mathbf{H}}_1}{\sqrt{n}} \xrightarrow{w.p.1} \rho_1 \mathbf{I}, \tag{6.49}$$

where $\rho_1 = \frac{\pi\sigma_s\sigma_d}{8M}\sqrt{\frac{p\eta}{\sigma_s^2 P_s+1}}$. The auto-covariance of equivalent noise $\tilde{\mathbf{w}}_1$ is

$$\frac{1}{n}\mathbf{W}_1 \xrightarrow{w.p.1} \lambda_1 \mathbf{I}, \tag{6.50}$$

where $\lambda_1 = \frac{\eta\sigma_d^2}{2(\sigma_s^2 P_s+1)} + 1$. Therefore, the end-to-end channel is diagonalized for large n, and similar to the results in Theorem 6.4.2, the rate achieved $R^{(1)}$ during frame 2 satisfies:

$$R^{(1)} - M\log\left(1 + \frac{np\pi^2\sigma_s^2\sigma_d^2\eta P_s}{32M^3(\eta\sigma_d^2 + 2\sigma_s^2 P_s + 2)}\right) \xrightarrow{w.p.1} 0. \tag{6.51}$$

Rate Achieved by $\mathcal{G}_2$

During frame 2, the relays in $\mathcal{G}_2$ receive the signal vector:

$$\mathbf{r}_2 = \sqrt{\frac{P_s}{M}}\mathbf{H}_2\mathbf{s}_2 + \mathbf{H}_r\mathbf{D}_1\left(\sqrt{\frac{P_s}{M}}\mathbf{H}_1\mathbf{s}_1 + \mathbf{n}_1\right) + \mathbf{n}_2, \tag{6.52}$$

where $\mathbf{s}_2$ is the signal sent by the source during frame 2, and the second term corresponds to the interference from the transmission of $\mathcal{G}_1$. During frame 3 the relays in $\mathcal{G}_2$ transmit to the destination with a processing matrix

$$\mathbf{D}_2 = \mathrm{diag}(T_{\frac{n}{2}+1}\, c_{\frac{n}{2}+1}, \cdots, T_n c_n), \tag{6.53}$$

where, to satisfy the power constraints, for $\frac{n}{2}+1 \le i \le n$

$$c_i = e^{j\theta_i}\sqrt{\frac{\eta}{np(P_s\sigma_s^2 + \eta\sigma_r^2/2 + 1)}}. \tag{6.54}$$

At the end of frame 3, the received signal at the destination is

$$\begin{aligned}\mathbf{y}_2 =& \mathbf{F}_2\mathbf{D}_2\mathbf{r}_2 + \mathbf{w}_2 \\ =& \sqrt{\frac{P_s}{M}}\underbrace{\mathbf{F}_2\mathbf{D}_2\mathbf{H}_2}_{\tilde{\mathbf{H}}_2}\mathbf{s}_2 + \sqrt{\frac{P_s}{M}}\mathbf{F}_2\mathbf{D}_2\mathbf{H}_r\mathbf{D}_1\mathbf{H}_1\mathbf{s}_1 + \underbrace{\mathbf{F}_2\mathbf{D}_2\mathbf{H}_r\mathbf{D}_1\mathbf{n}_1 + \mathbf{F}_2\mathbf{D}_2\mathbf{n}_2 + \mathbf{w}_2}_{\tilde{\mathbf{w}}_2} \quad .\end{aligned} \tag{6.55}$$

After correctly decoding $\mathbf{s}_1$, the destination cancels the inter-relay interference,[3] i.e., the second term in (6.55). After eliminating the inter-relay interference, we have an equivalent channel:

$$\mathbf{y}_2 = \sqrt{\frac{P_s}{M}}\tilde{\mathbf{H}}_2\mathbf{s}_2 + \tilde{\mathbf{w}}_2. \tag{6.56}$$

Following steps similar to (6.21) and (6.22), we have

$$\frac{\tilde{\mathbf{H}}_2}{\sqrt{n}} \xrightarrow{w.p.1} \rho_2\mathbf{I}, \tag{6.57}$$

where $\rho_2 = \frac{\pi\sigma_s\sigma_d}{8M}\sqrt{\frac{p\eta}{\sigma_s^2P_s+\eta\sigma_r^2/2+1}}$. Note that $\tilde{\mathbf{w}}_2$ is still a zero-mean Gaussian vector with auto-covariance

$$\frac{1}{n}\mathbf{W}_2 = \mathbf{F}_2\mathbf{D}_2\mathbf{H}_r\mathbf{D}_1\mathbf{D}_1^{\dagger}\mathbf{H}_r^{\dagger}\mathbf{D}_2^{\dagger}\mathbf{F}_2^{\dagger} + \mathbf{F}_2\mathbf{D}_2\mathbf{D}_2^{\dagger}\mathbf{F}_2^{\dagger} + \mathbf{I}. \tag{6.58}$$

In the right hand side of the above equation, we have

$$\begin{aligned}\mathbf{H}_r\mathbf{D}_1\mathbf{D}_1^{\dagger}\mathbf{H}_r^{\dagger} &= \frac{\eta}{np(P_s\sigma_s^2+1)}\mathbf{H}_r \operatorname{diag}(T_1,\cdots,T_{\frac{n}{2}})\,\mathbf{H}_r^{\dagger} \\ &\xrightarrow{w.p.1} \frac{\eta\sigma_r^2}{2(P_s\sigma_s^2+1)}\mathbf{I}.\end{aligned} \tag{6.59}$$

Therefore,

$$\mathbf{W}_2 \xrightarrow{w.p.1} \lambda_2\mathbf{I}, \tag{6.60}$$

where

$$\lambda_2 = \frac{1}{2(P_s\sigma_s^2+\eta\sigma_r^2/2+1)}\left[\frac{\eta^2\sigma_d^2\sigma_r^2}{2(P_s\sigma_s^2+1)} + \eta\sigma_d^2\right] + 1. \tag{6.61}$$

Combining (6.57) and (6.58) , the rate achieved by $\mathcal{G}_2$ is $R^{(2)}$ where

$$R^{(2)} - M\log\big(1+\frac{nP_s\rho_2^2}{M\lambda_2}\big) \xrightarrow{w.p.1} 0. \tag{6.62}$$

The overall rate is given by the following theorem.

[3] Interference cancellation requires knowledge of $\mathbf{H}_r$ at the destination, however, we note that even without this knowledge it is possible to obtain the same scaling of secondary throughput with the number of relays. Intuitively, the inter-relay interference is bounded by a constant that is under our control.

Theorem 6.6.1 *Consider a secondary system with an M-antenna source, an M-antenna destination, and n single-antenna relays, in the presence of N primary nodes each tolerating interference no more than γ. The secondary rate satisfies*

$$\overline{R} - \left(R^{(1)} + R^{(2)}\right)/3 \xrightarrow{w.p.1} 0, \tag{6.63}$$

under the Alternating Relaying Protocol with $L = 3$ and fixed clustering.

From Theorem 6.6.1, the growth rate of $\overline{R}$ is

$$\frac{\overline{R}}{\frac{2M}{3}\log n} \xrightarrow{w.p.1} 1. \tag{6.64}$$

Remark 6.6.2 *Theorem 6.6.1 can be generalized to an arbitrary number of transmission blocks L. For general L we can conclude:*

$$\frac{\overline{R}}{\frac{(L-1)M}{L}\log n} \xrightarrow{w.p.1} 1.$$

As L increases, the growth rate of $\overline{R}$ approaches the maximum value of $M\log n$.

6.7 Numerical Results

Unless otherwise specified, we use parameters $\bar{P}_s = 10$, $M = 2$, $N = 1$, $\gamma = 5$ and $\sigma_s^2 = \sigma_d^2 = 1$.

The secondary rates as a function of source transmit power are presented by Figure 6.5. The theoretical rate under various P_s is calculated according to (6.30) and (6.34). Recall that the theoretically optimal P_s given by (6.40) is obtained by (6.30) and (6.34). When the source interference links are very weak, e.g., $\sigma_{sp}^2 = 0.1$, maximizing the source power is optimal, which is similar to non-spectrum-sharing networks. When the source interference links is strong, e.g., $\sigma_{sp}^2 = 1, 2$, unlike non-spectrum-sharing networks, the secondary achieves higher rate if the source transmit at power lower than the maximum value. This is because the source needs to ensure the relays can operate with sufficient power, subject to the total interference constraints on the primary nodes.

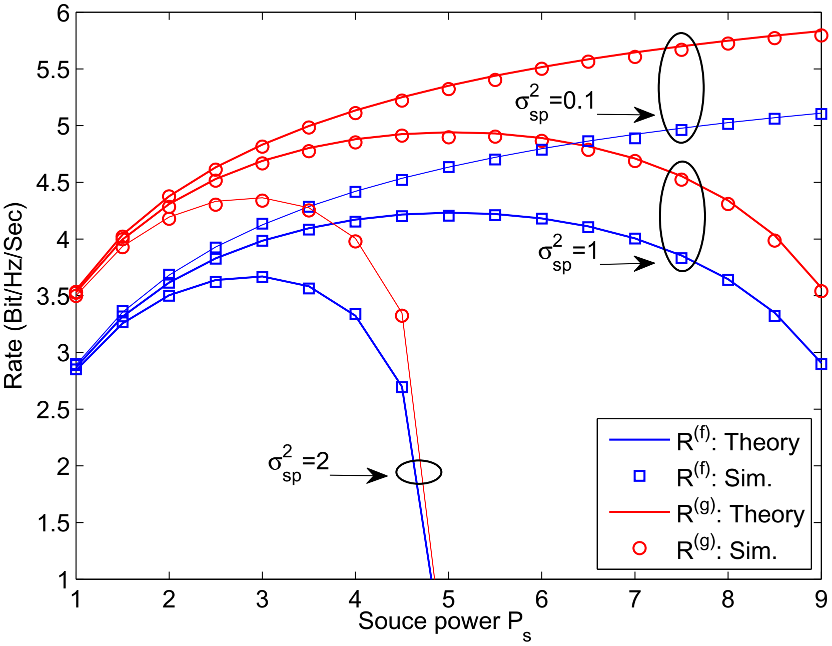

Figure 6.5. Throughput as a function of source power when $n = 100, \sigma_2 = \sigma_d = \sigma_p = 1$

Figure 6.6 verifies Theorem 6.4.2 under fixed clustering and gain clustering. Here, $\alpha = \gamma$, $\sigma_{sp}^2 = 1$, P_r is given by (6.35) and $P_s = 5$, which is almost optimal as shown in Figure 6.5. The simulated average rate of R_n under two clustering schemes are compared to $R^{(f)}$ given by (6.30) and $R^{(g)}$ given by (6.34), respectively, where the results are well matched for modest value of n. The secondary rate increases as the interference links of relays become weaker (smaller σ_p^2), since the relays can transmit at higher power (but the sum relay power is still bounded with n).

Figure 6.7 illustrates the tradeoff between maximizing secondary rate and minimizing interference on the primary. The interference power is $\gamma = 5(n)^{-\delta}$ with $\delta = 0.1$ and 0.2, respectively. For $\delta = 0.2$, the interference power decreases faster than $\delta = 0.1$, while the secondary rate increases more slowly.

The rate of Alternating Relaying Protocol (Theorem 6.6.1) is shown in Figure 6.8. We consider $\alpha \to \infty$, where all the relays in $\mathcal{G}_1$ and $\mathcal{G}_2$ transmit alternatively. Here, $P_s = 5$

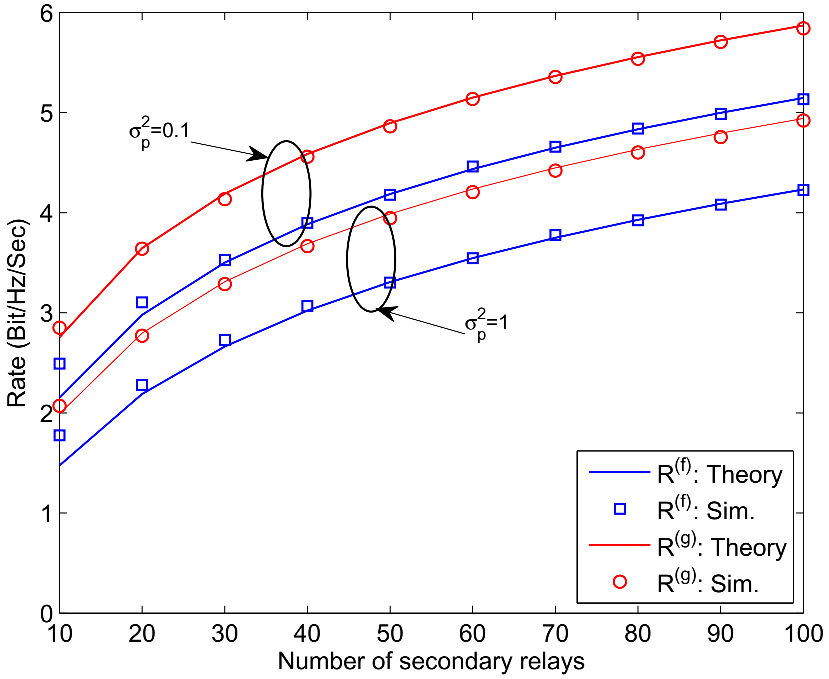

Figure 6.6. Secondary rate under two clustering schemes

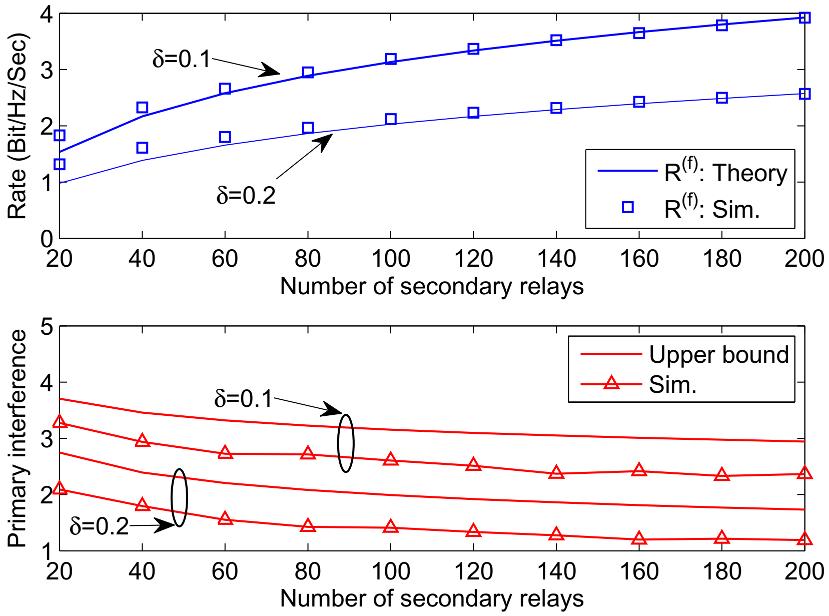

Figure 6.7. Secondary rate and primary interference as a function of number of relays

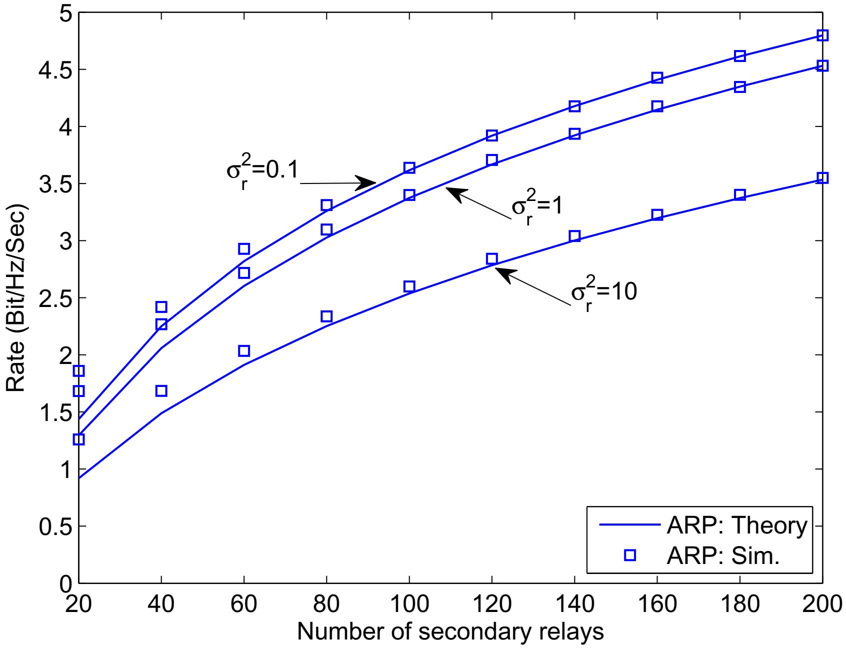

Figure 6.8. Secondary rate under the alternating relaying protocol

and P_r is determined by (6.45). The simulated rates match the theoretic analysis well under modest value of n. As the relay-relay channel becomes weaker (smaller σ_r^2), the inter-relay interference is reduced, and thus the secondary rate increases.

CHAPTER 7
CONCLUSIONS AND FUTURE WORK

This chapter summarizes the contributions of this dissertation and provides some possible avenues for future research. The results of this dissertation appear in [30, 52, 53, 67–77].

In the first part of this dissertation, we propose a product superposition signaling that significantly improves the rate performance of the MIMO broadcast channel with varying CSIR. First, two product superposition methods based on Grassmannian signaling are analyzed, and are shown to attain higher DoF than TDMA and indeed are optimal for a wide set of antenna configuration and channel coherence time. Then, we extend the product superposition to coherent signaling. The proposed method transmits a product of two signal matrices for the static and dynamic user, respectively, and each user decodes its own signal in a conventional manner. The method can work without interference cancellation, therefore has low complexity. For the entire SNR range, the static user attains considerable rate almost without degrading the rate for the dynamic user. The static user's rate is further improved by allowing the static user to cancel the dynamic user's signal.

It is possible to extend the above-mentioned results to more than two receivers. The set of receivers can be divided into two sets, one of them the dynamic set and the other the static set. At each point in time, the transmitter uses product superposition to broadcast to two users, one from each group. A scheduler selects the pair of users that is serviced at each time. To facilitate the case where there are unequal number of dynamic and static users, the pair memberships are allowed to be non-unique, i.e., there may be two or more pairs that contain a given receiver.

Note that throughout this work, both users are assumed to be in an ergodic mode of operation, i.e., the codewords are sufficiently long to allow coding arguments to apply. Simple extensions to this setup are easily obtained. For example, if the static user's coherence time is very long, one may adapt the transmission rate of the static user to its channel but allow the dynamic user to remain in an ergodic mode. Most expressions in this paper remain the same, except that for the rates and powers of the static user, expected values will be replaced with constant values.

In second part of this dissertation, we study the performance limits of an underlay cognitive network consisting of a multi-user and multi-antenna primary and secondary systems. We find the throughput limits of the secondary system as well as the tradeoff between this throughput and the tightness of constraints imposed by the primary system. Given a set of interference power constraints on the primary, the maximum throughput of the secondary MAC grows as $\frac{m}{N_p+1}\log n$ (primary broadcast), and $\frac{m}{M_p+1}\log n$ (primary MAC). These growth rates are attained by a simple threshold-based user selection rule. Interestingly, the secondary system can force its interference on the primary to zero while maintaining a growth rate of $\Theta(\log n)$. For the secondary broadcast channel, the secondary throughput can grow as $m \log\log n$ in the presence of either the primary broadcast or MAC channel. The growth rate of the throughput is unaffected by the presence of the primary (thus optimal). Furthermore, the interference on the primary can also be made to decline to zero, while maintaining the secondary throughput to grow as $\Theta(\log\log n)$.

Furthermore, we propose a Hybrid Opportunistic Scheduling for cognitive MAC, which is driven by two objectives: Maximizing the secondary throughput and minimizing the primary interference. The proposed scheme strictly controls the primary interference by opportunistic interference avoidance, and enhances the secondary throughput by activating transmitters with large secondary-channel gain. We characterize the optimal number of active secondary transmitters and a tradeoff between the secondary throughput enhancement and the primary interference reduction. Finally, we study user scheduling under a fairness constraint when links have non-i.i.d. statistics.

Finally, we study spectrum sharing networks with distributed amplify-forward relaying to improve the secondary rate and reduce the interference on the primary. In the asymptote of large n (number of relays) the optimal power strategy for the secondary source and relays was found, achieving a secondary rate proportionally to $\log n$. The half-duplex rate loss was reduced and the scaling of secondary rate was enhanced by the introduction of the Alternating Relay Protocol. The trade-off between the secondary rate and the interference on the primary was characterized. Finally, our results show that even without cross channel information at the secondary, the secondary rate can achieve the growth rate $\log n$.

REFERENCES

[1] Federal Communications Commission, “Facilitating opportunities for flexible, efficient, and reliable spectrum use employing cognitive radio technologies,” Dec. 2003.

[2] A. Özgür, O. Lévêque, and D. Tse, “Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks,” *IEEE Trans. Inform. Theory*, vol. 53, no. 10, pp. 3549 –3572, Oct. 2007.

[3] Y. Liang and V. V. Veeravalli, “Cooperative relay broadcast channels,” *IEEE Trans. Inform. Theory*, vol. 53, no. 3, pp. 900 –928, Mar. 2007.

[4] D. Chen, K. Azarian, and J. Laneman, “A case for amplify-forward relaying in the block-fading multiple-access channel,” *IEEE Trans. Inform. Theory*, vol. 54, no. 8, pp. 3728 –3733, Aug. 2008.

[5] O. Şahin, O. Simeone, and E. Erkip, “Interference channel with an out-of-band relay,” *IEEE Trans. Inform. Theory*, vol. 57, no. 5, pp. 2746 –2764, May 2011.

[6] G. Zhao, J. Ma, G. Li, T. Wu, Y. Kwon, A. Soong, and C. Yang, “Spatial spectrum holes for cognitive radio with relay-assisted directional transmission,” *IEEE Trans. Wireless Commun.*, vol. 8, no. 10, pp. 5270 –5279, Oct. 2009.

[7] J. Mietzner, L. Lampe, and R. Schober, “Distributed transmit power allocation for multihop cognitive-radio systems,” *IEEE Trans. Wireless Commun.*, vol. 8, no. 10, pp. 5187 –5201, Oct. 2009.

[8] V. Asghari and S. Aissa, “Cooperative relay communication performance under spectrum-sharing resource requirements,” in *Proc. IEEE ICC*, May 2010.

[9] L. Li, X. Zhou, H. Xu, G. Li, D. Wang, and A. Soong, “Simplified relay selection and power allocation in cooperative cognitive radio systems,” *IEEE Trans. Wireless Commun.*, vol. 10, no. 1, pp. 33 –36, Jan. 2011.

[10] M. Naeem, D. Lee, and U. Pareek, “An efficient multiple relay selection scheme for cognitive radio systems,” in *Proc. IEEE ICC*, May 2010.

[11] Y. Zou, J. Zhu, B. Zheng, and Y.-D. Yao, “An adaptive cooperation diversity scheme with best-relay selection in cognitive radio networks,” *IEEE Trans. Signal Processing*, vol. 58, no. 10, pp. 5438 –5445, Oct. 2010.

[12] J. Lee, H. Wang, J. Andrews, and D. Hong, “Outage probability of cognitive relay networks with interference constraints,” *IEEE Trans. Wireless Commun.*, vol. 10, no. 2, pp. 390 –395, Feb. 2011.

[13] G. Caire and S. Shamai, “On the achievable throughput of a multiantenna Gaussian broadcast channel,” *IEEE Trans. Inform. Theory*, vol. 49, no. 7, pp. 1691 – 1706, July 2003.

[14] C. Huang, S. Jafar, S. Shamai, and S. Vishwanath, “On degrees of freedom region of MIMO networks without channel state information at transmitters,” *IEEE Trans. Inform. Theory*, vol. 58, no. 2, pp. 849 –857, Feb. 2012.

[15] S. Jafar and A. Goldsmith, “Isotropic fading vector broadcast channels:the scalar upper bound and loss in degrees of freedom,” *IEEE Trans. Inform. Theory*, vol. 51, no. 3, pp. 848 – 857, Mar. 2005.

[16] T. S. Rappaport, *Wireless Communications: Principles and Practice.* Prentice Hall PTR, 1995.

[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* John Wiley and Sons, 1991.

[18] A. El Gamal, “The capacity of a class of broadcast channels,” *IEEE Trans. Inform. Theory*, vol. 25, no. 2, pp. 166 – 169, 1979.

[19] J. Korner and K. Marton, “General broadcast channels with degraded message sets,” *IEEE Trans. Inform. Theory*, vol. 23, no. 1, pp. 60 – 64, 1977.

[20] L. Zheng and D. N. C. Tse, “Communication on the Grassmann manifold: a geometric approach to the noncoherent multiple-antenna channel,” *IEEE Trans. Inform. Theory*, vol. 48, no. 2, pp. 359–383, 2002.

[21] T. L. Marzetta and B. M. Hochwald, “Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading,” *IEEE Trans. Inform. Theory*, vol. 45, no. 1, pp. 139–157, 1999.

[22] W. M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry.* Academic Press, 1986.

[23] J. G. Proakis, *Digital Communications.* McGraw-Hill, 2001.

[24] B. M. Hochwald and T. L. Marzetta, “Unitary space-time modulation for multiple-antenna communications,” *IEEE Trans. Inform. Theory*, vol. 46, no. 2, pp. 543–564, 2000.

[25] B. Hassibi and B. M. Hochwald, “How much training is needed in multiple-antenna wireless links?” *IEEE Trans. Inform. Theory*, vol. 49, no. 4, pp. 951–963, 2003.

[26] E. Telatar, “Capacity of multi-antenna Gaussian channels,” *Euro. Trans. on Telecomm.*, vol. 10, no. 6, pp. 585–595, 1999.

[27] A. Edelman, T. A. Arias, and S. T. Smith, “The geometry of algorithms with orthonormality constraints,” *Appl. Comput. Harmonic Anal.*, vol. 20, no. 2, pp. 303–353, 1998.

[28] R. J. Muirhead, *Aspects of Multivariate Statistical Theory.* Wiley, 1982.

[29] R. A. Horn and C. R. Johnson, *Matrix Analysis.* Cambridge University Press, 1985.

[30] Y. Li and A. Nosratinia, “Product superposition for MIMO broadcast channel,” *IEEE Trans. Inform. Theory*, vol. 58, no. 11, pp. 6839 –6852, Nov. 2012.

[31] H. Weingarten, Y. Steinberg, and S. Shamai, “The capacity region of the Gaussian multiple-input multiple-output broadcast channel,” *IEEE Trans. Inform. Theory*, vol. 52, no. 9, pp. 3936 –3964, Sept. 2006.

[32] N. Jindal, “MIMO broadcast channels with finite-rate feedback,” *IEEE Trans. Inform. Theory*, vol. 52, no. 11, pp. 5045 –5060, Nov. 2006.

[33] T. Yoo and A. Goldsmith, “On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming,” *IEEE J. Select. Areas Commun.*, vol. 24, no. 3, pp. 528 –514, Mar. 2006.

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

[35] A. Ghasemi and E. S. Sousa, “Fundamental limits of spectrum-sharing in fading environments,” *IEEE Trans. Wireless Commun.*, vol. 6, no. 2, pp. 649–658, Feb. 2007.

[36] R. Zhang and Y.-C. Liang, “Exploiting multi-antennas for opportunistic spectrum sharing in cognitive radio networks,” *IEEE J. Select. Topics Signal Processing*, vol. 2, no. 1, pp. 88 –102, Feb. 2008.

[37] R. Zhang, S. Cui, and Y.-C. Liang, “On ergodic sum capacity of fading cognitive multiple-access and broadcast channels,” *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 5161–5178, Nov. 2009.

[38] M. Gastpar, “On capacity under receive and spatial spectrum-sharing constraints,” *IEEE Trans. Inform. Theory*, vol. 53, no. 2, pp. 471–487, Feb. 2007.

[39] P. Viswanath, D. N. C. Tse, and R. Laroia, “Opportunistic beamforming using dumb antennas,” *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1277–1294, June 2002.

[40] K. Hamdi, W. Zhang, and K. B. Letaief, "Opportunistic spectrum sharing in cognitive MIMO wireless networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 8, pp. 4098–4109, Aug. 2009.

[41] N. Jamal, H. E. Saffar, and P. Mitran, "Throughput enhancements in point-to-multipoint cognitive systems," in *Proc. IEEE ISIT*, June/July 2009, pp. 2742–2746.

[42] ——, "Asymptotic scheduling gains in point-to-multipoint cognitive networks." [Online]. Available: http://arxiv.org/pdf/1001.3365

[43] R. J. Serfling, *Approximation Theorems of Mathematical Statistics.* New York: Wiley, 1980.

[44] A. Moustakas, S. Simon, and A. Sengupta, "MIMO capacity through correlated channels in the presence of correlated interferers and noise: a (not so) large n analysis," *IEEE Trans. Inform. Theory*, vol. 49, no. 10, pp. 2545 – 2561, Oct. 2003.

[45] H. Bolcskei, R. Nabar, O. Oyman, and A. Paulraj, "Capacity scaling laws in MIMO relay networks," *IEEE Trans. Wireless Commun.*, vol. 5, no. 6, pp. 1433 –1444, June 2006.

[46] S. Diggavi and T. Cover, "The worst additive noise under a covariance constraint," *IEEE Trans. Inform. Theory*, vol. 47, no. 7, pp. 3072 –3081, Nov. 2001.

[47] B. Hochwald, T. Marzetta, and V. Tarokh, "Multiple-antenna channel hardening and its implications for rate feedback and scheduling," *IEEE Trans. Inform. Theory*, vol. 50, no. 9, pp. 1893 – 1909, Sept. 2004.

[48] O. Oyman, R. Nabar, H. Bolcskei, and A. Paulraj, "Characterizing the statistical properties of mutual information in MIMO channels," *IEEE Trans. Signal Processing*, vol. 51, no. 11, pp. 2784 – 2795, Nov. 2003.

[49] H. A. David and H. N. Nagaraja, *Order statistics.* Wiley, 2003.

[50] N. T. Uzgoren, "The asymptotic development of the distribution of the extreme values of a sample," in *Studies in Mathematics and Mechanics: Presented to Richard von Mises by Friends, Colleagues, and Pupils.* New York: Academic, 1954, pp. 346–353.

[51] A. Tajer and X. Wang, "Multiuser diversity gain in cognitive networks with distributed spectrum access," in *Proc. Information Sciences and Systems (CISS)*, Mar. 2009, pp. 135 –140.

[52] Y. Li and A. Nosratinia, "Capacity limits of multiuser multiantenna cognitive networks," *IEEE Trans. Inform. Theory*, vol. 58, no. 7, pp. 4493 –4508, July 2012.

[53] ——, “Opportunistic cognitive radio broadcast channel: Asymptotic performance,” in *Globecom Workshop on Mobile Computing and Emerging Communication Networks*, Dec. 2010.

[54] C. Shen and M. Fitz, “Opportunistic spatial orthogonalization and its application in fading cognitive radio networks,” *IEEE J. Select. Topics Signal Processing*, vol. 5, no. 1, pp. 182–189, Feb. 2011.

[55] J.-P. Hong and W. Choi, “Capacity scaling law by multiuser diversity in cognitive radio systems,” in *Proc. IEEE ISIT*, June 2010, pp. 2088 –2092.

[56] R. Zhang and Y.-C. Liang, “Investigation on multiuser diversity in spectrum sharing based cognitive radio networks,” *IEEE Commun. Lett.*, vol. 14, no. 2, pp. 133 –135, Feb. 2010.

[57] T. W. Ban, W. Choi, B. C. Jung, and D. K. Sung, “Multi-user diversity in a spectrum sharing system,” *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 102 –106, Jan. 2009.

[58] S. Jayaweera and T. Li, “Dynamic spectrum leasing in cognitive radio networks via primary-secondary user power control games,” *IEEE Trans. Wireless Commun.*, vol. 8, no. 6, pp. 3300 –3310, June 2009.

[59] X. Liu, E. K. P. Chong, and N. B. Shroff, “A framework for opportunistic scheduling in wireless networks,” *Computer Networks*, vol. 41, no. 4, pp. 451 – 474, 2003.

[60] S. Patil and G. De Veciana, “Measurement-based opportunistic scheduling for heterogenous wireless systems,” *IEEE Trans. Commun.*, vol. 57, no. 9, pp. 2745 –2753, Sept. 2009.

[61] S. Sanayei and A. Nosratinia, “Opportunistic downlink transmission with limited feedback,” *IEEE Trans. Inform. Theory*, vol. 53, no. 11, pp. 4363–4372, Nov. 2007.

[62] A. Scaglione and Y.-W. Hong, “Opportunistic large arrays: Cooperative transmission in wireless multihop ad hoc networks to reach far distances,” *IEEE Trans. Signal Processing*, vol. 51, no. 8, pp. 2082 – 2092, Aug. 2003.

[63] R. Zhang, “On peak versus average interference power constraints for protecting primary users in cognitive radio networks,” *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 2112 –2120, Apr. 2009.

[64] Y. Jing and B. Hassibi, “Distributed space-time coding in wireless relay networks,” *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, pp. 3524 –3536, Dec. 2006.

[65] P. Gupta and P. Kumar, “The capacity of wireless networks,” *IEEE Trans. Inform. Theory*, vol. 46, no. 2, pp. 388 –404, Mar. 2000.

[66] A. Gut, *Stopped Random Walks: Limit Theorems and Applications.* Springer, 2008.

[67] Y. Li and A. Nosratinia, "Coherent product superposition for downlink multiuser MIMO," *IEEE Trans. Wireless Commun.*, submitted.

[68] ——, "Grassmannian-Euclidean superposition for MIMO broadcast channels," in *Proc. IEEE ISIT*, July 2012, pp. 2491 –2495.

[69] ——, "Broadcasting on the Grassmannian: Enhancing the multiplexing gain," in *Proc. IEEE ISIT*, Aug. 2011, pp. 1733 –1737.

[70] ——, "Spectrum sharing with distributed relay selection and clustering," *IEEE Trans. Commun.*, to appear.

[71] ——, "Hybrid opportunistic scheduling in cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 1, pp. 328 –337, Jan. 2012.

[72] ——, "Asymptotically optimal scheduling in cognitive multiple-access channel," in *Information Theory and Applications Workshop, IEEE*, Feb. 2011.

[73] ——, "Spectrum-sharing capacity enhancement with distributed relaying," in *Proc. IEEE ICC*, June 2012.

[74] ——, "Throughput limits and multiuser diversity of multiantenna spectrum sharing networks," in *Proc. IEEE GLOBECOM*, Dec. 2011.

[75] Y. Li, A. Nosratinia, and W. Zhang, "Opportunistic cooperation for distributed spectrum sensing in cognitive radio," in *Proc. IEEE ICC*, June 2011.

[76] B. Xie, Y. Li, H. Minn, and A. Nosratinia, "Adaptive interference alignment with CSI uncertainty," *IEEE Trans. Commun.*, to appear.

[77] ——, "Throughput limits and multiuser diversity of multiantenna spectrum sharing networks," in *Proc. IEEE GLOBECOM*, Dec. 2011.

VITA

Yang Li received a B.S. and a M.S. degree, in 2005 and 2008, in electronic engineering from Shanghai Jiao Tong University, Shanghai, China. He is currently a Ph.D. candidate in electrical engineering of The University of Texas at Dallas, Richardson, TX, USA. Since 2012, he has been with the Dallas Technology Lab (DTL) of Samsung Telecommunications America (STA), Texas, as a Senior Algorithms and Standards Engineer, focusing on the next generation wireless communications technologies and 3GPP Long Term Evolution-Advanced (LTE-A) Standard. He had intern positions at Huawei Technologies as a product manager and system engineer in 2011 and 2008, respectively. He has served as a technical reviewer for numerous journals and conferences, has been a student member of IEEE since 2009, and was the student lead of Chinese Institute of Engineer (CIE), Dallas Fort Worth Chapter in 2010. He is a member of Golden Key International Honor Society.