The Dissertation Committee for Jubin Jose
certifies that this is the approved version of the following dissertation:

Source and Channel Aware Resource Allocation for Wireless Networks

Committee:

Sriram Vishwanath, Supervisor

Jeffrey G. Andrews

Sanjay Shakkottai

Gustavo de Veciana

David Morton

Source and Channel Aware Resource Allocation for Wireless Networks

by

Jubin Jose, B. Tech.; M. S. E.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2011

To my parents, sister

and Rangu

Acknowledgments

My PhD has been an exciting, challenging, rewarding and enjoyable expedition. Throughout this journey, my advisor, Sriram Vishwanath, has been the primary source of encouragement and support. He introduced me to a wide range of areas, provided me with research directions, and supported me in pursuing the topics of my interest. While pursuing my interests, I greatly benefited from his profound knowledge, creative thinking, constructive criticism and most importantly optimism. Moreover, he created wonderful collaboration opportunities, encouraged me to participate in many professional activities, and thus, has been pivotal in my overall development as a researcher. I am extremely thankful to him for being such a brilliant mentor.

I also thank the rest of my PhD committee members: Jeff Andrews, Sanjay Shakkottai, Gustavo de Veciana and Dave Morton. Their willingness to help made them easily approachable. In this regard, I thank Jeff Andrews for always finding time for me even with the busy schedule of a WNCG director and for giving me insightful comments. I am grateful to Sanjay Shakkottai for providing constructive criticism on my research; Gustavo de Veciana for motivating me to think about intuitive explanations for even highly complicated problems and Dave Morton for his insightful questions during my presentations. I appreciate their willingness to serve on my PhD committee and

provide valuable feedback that improved the quality of my work. While taking graduate courses with them, I not only gained knowledge about the subject but also got inspired by their teaching skills. I would like to extend my gratitude to all my professors who have instilled in me a passion for learning and teaching.

As part of my dissertation work, I was fortunate to collaborate with some of the finest researchers in the field. In this regard, I would like to express my sincere gratitude to Lei Ying at Iowa State University, Narayan Prasad at NEC Labs America, and Alexei Ashikhmin and Tom Marzetta at Bell Labs. Working with them broadened my knowledge and provided me different perspectives towards solving a problem.

It goes without saying that my journey as a graduate student became more enjoyable by the company of my fellow group members. They have been very friendly and supportive throughout. I learnt a lot through the weekly group meetings, occasional study sessions and impromptu research discussions. Brian Smith, Caleb Lo, Sriram Sridharan and Shreeshankar Bodas aided me in improving my presentation skills. I greatly benefited from the company of Rajiv Soundararajan, Amin Jafarian and Goochul Chung both at UT and at conferences. I thoroughly enjoyed the numerous conversations, both technical and non-technical, I had with Kumar Appaiah, Ioannis Mitliagkas, Abhik Das, Shweta Agrawal, Sharayu Moharir, Sang Hyun Lee, Ahmed Abdel-Hadi, Fabio Fernandes, Vidur Bhargava, Ankit Singh Rawat, Deepjyoti Deka, O. Ozan Koyluoglu, Kannan Srinivasan, Yongseok Yoo, Hongbo Si, Cong Li and

others in my group. I would like to take this opportunity to thank each one of them.

WNCG is literally the place where I spent majority of my time doing research. I consider myself lucky to be part of this premier research facility that made my graduate life both productive and fun at the same time. Research discussions and assistance were never limited by group boundaries in WNCG. I am thankful to Sundar Subramanian, Aneesh Reddy, Siddhartha Banerjee, Harish Ganapathy, Aditya Gopalan and all my colleagues for creating an environment that is conducive to learning. Without the valuable service of the WNCG staff - Janet Preuss, Julie Levy, Jennifer Graham and Adrian Duran - life at WNCG would not have been so smooth. Janet and Jennifer deserve my special thanks for being exceedingly helpful. I also appreciate the assistance from ECE staff especially Melanie Gulick who was always ready to help with a smile.

I thank JP's Java and Mozart's Coffee Roasters where I spent quite a bit of time writing portions of my dissertation while sipping their freshly brewed coffee.

My life in Austin was made wonderful by all my friends - Aneesh, Hari, Ramya, Priyamvada, Yagna, Ashwini, Dharma, Reeja, Aji, Chinmayi and many more including those outside Austin. My family - parents, sister and brother-in-law - has been very supportive throughout my PhD. My parents have done everything to provide me the best education possible and have always encouraged me to pursue my interests. I am indebted to my parents

and my sister for their unconditional support and love.

Last, but not the least, I have been fortunate to be in the company of someone special who has inspired me to a better person. For this, I am extremely grateful to Rangu for her encouragement, support and love.

Source and Channel Aware Resource Allocation for Wireless Networks

Publication No. _____

Jubin Jose, Ph.D.
The University of Texas at Austin, 2011

Supervisor: Sriram Vishwanath

Wireless networks promise ubiquitous communication, and thus facilitate an array of applications that positively impact human life. At a fundamental level, these networks deal with compression and transmission of sources over channels. Thus, accomplishing this task efficiently is the primary challenge shared by these applications. In practice, sources include data and video while channels include interference and relay networks. Hence, effective source and channel aware resource allocation for these scenarios would result in a comprehensive solution applicable to real-world networks.

This dissertation studies the problem of source and channel aware resource allocation in certain scenarios. A framework for network resource allocation that stems from rate-distortion theory is presented. Then, an optimal decomposition into an application-layer compression control, a transport-layer congestion control and a network-layer scheduling is obtained. After deducing

insights into compression and congestion control, the scheduling problem is explored in two cross-layer scenarios. First, appropriate queue architecture for cooperative relay networks is presented, and throughput-optimality of network algorithms that do not assume channel-fading and input-queue distributions are established. Second, decentralized algorithms that perform rate allocation, which achieve the same overall throughput region as optimal centralized algorithms, are derived.

In network optimization, an underlying throughput region is assumed. Hence, improving this throughput region is the next logical step. This dissertation addresses this problem in the context of three significant classes of interference networks. First, degraded networks that capture highly correlated channels are explored, and the exact sum capacity of these networks is established. Next, multiple antenna networks in the presence of channel uncertainty are considered. For these networks, robust optimization problems that result from linear precoding are investigated, and efficient iterative algorithms are derived. Last, multi-cell time-division-duplex systems are studied in the context of corrupted channel estimates, and an efficient linear precoding to manage interference is developed.

Table of Contents

List of Figures

Chapter 1

Introduction

Communication networks are highly complex systems that are designed and developed through multi-disciplinary work ranging from device-physics to application-software. The salient engineering principle behind the synergetic coexistence of all these disciplines is the layered architecture leading to increasing levels of *abstractions*. For instance, the physical-layer interprets the communication medium as a point-to-point memoryless channel to transmit information. The network-layer simplifies the network as a graph consisting of bit-pipes to route packets, while the transport-layer uses the notion of flows to control rate or congestion in the network. All these abstractions are conceptually simple and moderately effective for wired data networks.

Today's networks, both wired and wireless, despite the fact that these are primarily designed for data and voice, transfer massive amounts of video traffic including real-time streaming. Traffic forecasts predict that the fraction of video content being transferred is steadily growing along with the tremendous growth of overall traffic. For example, Cisco Visual Networking Index: Forecast and Methodology, 2009-2014 [2] reports the following in its global video highlights:

> "Internet video is now over one-third of all consumer Internet traffic, and will approach 40 percent of consumer Internet traffic by the end of 2010, not including the amount of video exchanged through P2P file sharing.
>
> The sum of all forms of video (TV, video on demand, Internet, and P2P) will continue to exceed 91 percent of global consumer traffic by 2014. Internet video alone will account for 57 percent of all consumer Internet traffic in 2014."

Since video traffic is fundamentally different from data and voice, these forecasts necessitate the development of rich-enough abstractions suited for designing networks that carry video. This will enable the re-design of network architectures and protocols to account for and exploit the properties of the source (video). We call such an approach *source-aware* resource allocation.

While zooming in on *wireless* networks, the inherent distinction from wired networks is the underlying *shared* wireless medium for communication. Due to extensive research in wireless over the past few decades, researchers have somewhat mastered the art of reliably communicating over a point-to-point wireless channel. Despite all the research efforts, there is limited knowledge about the fundamental limits of wireless networks. This is due to the fact that the wireless medium is shared, which invalidates the simple abstraction of a graph consisting of bit-pipes. In other words, these wireless links interact (or interfere) in potentially complex ways through the wireless medium, which is typically captured through physical-layer models that involve all (if

not subsets of) nodes in the network. The maximum achievable data-rates under such sophisticated physical-layer models are studied in network information theory literature. While moderate advances have been made in this area, a complete characterization as in the point-to-point case [113] does not exist to date. Thus, understanding the maximum achievable data-rates of different channels, especially interference networks, is one of the crucial steps to design more efficient wireless communication systems. In real-world networks, the interference issue is worsening as a result of decreasing cell-sizes[1] along with aggressive frequency reuse. In parallel with studying optimal communication schemes, for immediate applicability in these real-world networks, it is beneficial to have low-complexity schemes that require limited channel state information (CSI). We call such an approach *channel-aware* resource allocation.

Source and channel aware resource allocation, like traditional[2] resource allocation, (adaptively) operates a network at different operating points based on the network state (e.g., channel conditions) and the load (i.e., backlog) in the network. There are well-established methodologies for designing such algorithms, but, in general, these techniques are not successful in two separate aspects. First, very little is known about applying existing techniques to cooperative networks. The main challenge in addressing this is the cou-

[1] In cellular networks, smaller cell-sizes result from increasing macro cell density and/or deployment of pico and femto cells.

[2] By traditional resource allocation, we refer to allocation of orthogonal resources such as time-slots and frequency-bands.

pling between cooperative achievable schemes and queueing, which are often studied separately. Next, there is limited knowledge in designing completely decentralized (or distributed) algorithms without compromising throughput, i.e., without reducing data-rates. Such algorithms are applicable in wireless local area networks (WLANs) and ad hoc networks that lack centralized infrastructure. Thus, overcoming these limitations, along with being source and channel aware, would result in a powerful end-to-end solution for addressing a majority of the challenges in ad hoc, cellular and relay (wireless) networks. This forms the underlying theme of this dissertation.

Next, we provide an overview of the specific problems studied in this dissertation. This can be logically subdivided into two (related) topics: (*i*) network architecture, and (*ii*) interference management.

1.1 Network Architecture: Overview

Majority of existing network architectures and protocols are optimized assuming pre-compressed and packetized data [110]. For video, this assumption of data packets abstracts out one of the key aspects - the lossy compression problem. Hence, in addition to source-rate, an abstraction of sources like video would at least require (average) *distortion* resulting from compression of sources. For such source abstractions, we develop a new framework for network resource allocation that stems from rate-distortion theory [10]. One of the main results obtained using this framework is the decomposition of the resource allocation problem into an application-layer compression control, a

transport-layer congestion control and a network-layer scheduling. We show that this decomposition is provably optimal for the class of concave objective functions. Then, we derive new insights from the developed rate-distortion framework by focusing on special cases, in particular, congestion control for binary sources and optimal resource allocation for multiple access channels.

In the above decomposition into layers, the component that is heavily dependent on network-type (e.g., infrastructure-less, centralized, cooperative) is the network-layer scheduling. The backpressure algorithm introduced in the context of stable operation of networks in [119] can be viewed as a dynamic solution to this problem. This algorithm and its extensions apply to a wide-range of networks including multi-hop networks and perform *throughput-optimal* scheduling (and routing). However, very little is known in applying this in the context of cooperative networks. The main challenge is that an abstraction of the physical layer coding using coupled data-rates is insufficient to capture notions such as node cooperation in cooperative relay networks. Consequently, network-stability analyses based on such abstractions are valid for non-cooperative schemes alone and meaningless for cooperative schemes.

Motivated from this, we develop cooperative queueing architectures that bring the information-theoretic (cooperative) coding scheme together with network-stability analysis. Our cooperative queueing framework does not constrain the system to any particular achievable scheme, i.e., the relays can use any cooperative coding strategy of its choice such as amplify, compress, quantize or any alter-and-forward scheme. We consider the demanding scenario

when coherence duration is of the same order of the packet/codeword duration, the channel distribution is unknown and the fading state is only known causally. For this scenario, our main contributions are two-fold: first, we develop a low-complexity queue-architecture to enable stable operation of cooperative relay networks, and, second, we establish the throughput optimality of a simple network algorithm that utilizes this queue-architecture.

So far, we have a network architecture with distributed[3] algorithms at the application-layer and the transport-layer. However, the network-layer scheduling is a centralized and often hard optimization problem. There is a lot of interest (and literature) in developing decentralized algorithms. Very recently, for networks with carrier sensing, a fully decentralized algorithm has been proposed in [59]. This decentralized queue-length based scheduling and its variants have been shown to be throughput-optimal in [57,58,81]. However, these results are limited to on-off scheduling. To optimally utilize physical-layer schemes for wireless networks (including WLANs and ad hoc networks), we need decentralized cross-layer algorithms that jointly perform medium access control (on-off scheduling) and physical-layer rate adaptation. Hence, we generalize the existing on-off framework to multi-state framework. As the main result, we provide a distributed rate allocation algorithm for wireless networks and establish that this algorithm is throughput-optimal for general rate regions. The algorithm requires that each link can determine the global feasi-

[3]Note that the different layers resulting from the decomposition are coupled through dual variables.

bility of increasing its current data-rate. We provide insights on approximately achieving this in practice.

In network resource allocation, an underlying throughput region for the network is assumed. Hence, improving this throughput region is part of the resource allocation problem itself. This is a highly non-trivial problem, and has to be addressed on a case-by-case basis. Next, we study this problem for important classes of interference networks.

1.2 Interference Management: Overview

The interest in the interference channel (IFC) and its fundamental limits stems from the wide range of applications that will benefit from such an analysis. However, large gaps exist in our understanding of interference channels. Since the introduction of interference channels [4], the class of two-transmitter two-receiver interference channels has been studied in great detail. Indeed, a majority of exact or approximate capacity results are known only for such two-user interference channels [7, 22, 29, 41, 88, 101, 103, 112]. However, real-world applications such as cellular networks involve many users. In contrast to two-user IFCs, three-or-more user IFCs are fundamentally different due to the presence of multiple interferes. The interfering signals present in the network can be managed in multiple ways, for example, by treating interference as noise, through interference cancellation and interference avoidance. However, the optimal scheme is an open problem even after four decades of research. Given the difficulty of finding the capacity region of interference

networks, it is logical to make progress by focusing on a class of channels or a class of achievable schemes.

We develop a family of *genie*-MAC (multiple access channel) outer bounds for Gaussian K-user interference channels. This family is based on existing genie-aided bounding mechanisms, but differs from current approaches in its optimization problem formulation and application. The genie-aided bound (based on [76]) creates a group of genie receivers that form MACs that can decode a subset of the original IFC's messages. The MAC sum-capacity of each of the genie receivers provides an outer bound on the sum of rates for this subset. Thus, the genie-aided MAC bound is formulated as an optimization problem. We show that this outer bound is tight in sum-capacity for the degraded Gaussian K-user IFC. *Degraded* networks is the class of interest to study extremely correlated channels. Thus, we establish[4] the sum-capacity of degraded Gaussian interference networks. This result holds true for any number of users and any signal-to-noise ratios (SNRs). The scheme that achieves sum-capacity is successive interference cancellation. We generalize this framework to unit-rank interference channels where each user is equipped with multiple antennas. Again, the sum-capacity of the MIMO (multiple input multiple output) unit-rank Gaussian IFC is characterized using a combination of a optimization problem resulting from a genie-aided outer bound and successive interference cancellation as the achievable scheme. The proof in the MIMO

[4]An alternate proof based on entropy power inequality (EPI) is possible. Our proof is interesting as it shows the tightness of genie-MAC outer bound for degraded channels.

case establishes a simple equivalence between the MIMO unit-rank K-user IFC and a corresponding SISO degraded K-user IFC.

There are very few exact capacity results known in literature for K-user Gaussian IFCs. However, significant progress has been made in recent literature by limiting results to asymptotic optimality, in particular, degrees-of-freedom optimality. Interference alignment has been shown to be degrees-of-freedom optimal for time varying (or frequency selective) interference channels including MIMO [18, 47]. These results motivate *linear precoding* used for alignment as a reasonable scheme for interference networks with multiple antennas. Additionally, such achievable schemes are greatly motivated from an implementation perspective in cellular standards such as LTE-A - CoMP: Coordinated Multi-Point TX/RX and 802.16m - Multi-BS MIMO. Further, these standards require distributed schemes based on limited (imperfect) channel state information (CSI) at the transmitters. Even though good[5] iterative linear precoding schemes have been developed under perfect CSI [45, 98, 105], there is limited knowledge with imperfect CSI.

Therefore, we study a robust weighted-sum rate optimization problem in the presence of *channel uncertainty* for *multiple antenna* Gaussian interference networks. In this robust formulation, receivers have perfect CSI while transmitters have imperfect CSI, and the resulting precoder choice will work for any channel realization within the uncertainty region. Unfortunately, due

[5]Linear precoding in its existing formulations are hard non-convex problems. Hence, many iterative schemes have been developed.

to the hardness of this problem, optimal solution cannot be efficiently obtained. Instead of resorting to ad-hoc algorithms, we show that it is possible to design algorithms using a systematic approach. Towards this end, we develop new provably convergent iterative algorithms for precoder design through ingenious sub-problem formulations such that each of these sub-problems can be solved optimally. The sub-problems are solved in closed-form for certain cases and formulated as standard convex problems for the rest. Next, this approach is extended to robust max-min rate optimization. Finally, to complement these contributions on achievable schemes, we generalize the genie-MAC outer bounding technique to incorporate channel uncertainty using notions of compound-MAC capacity and then obtain computable outer bounds using an alternating optimization approach.

The robust formulation above assumes perfect CSI at the receivers.[6] We relax this assumption and jointly study channel estimation and linear precoding in time-division-duplex (TDD) systems. We consider a multi-cell multiple antenna system with precoding used at the base stations for downlink transmission. A popular technique for obtaining CSI in TDD systems is uplink training by utilizing the reciprocity of the wireless medium. We mathematically characterize the impact that uplink training has on the performance of such multi-cell TDD systems. When non-orthogonal training sequences are used for uplink training, we show that the precoding matrix used by the

[6]This is a widely used assumption in studying frequency-division-duplex (FDD) downlink systems. It helps to focus on the more crucial limited CSI aspect at the base-stations.

base station in one cell becomes corrupted by the channel between that base station and the users in other cells in an undesirable manner. We analyze this fundamental problem of *pilot contamination* in multi-cell systems. Next, we develop a new multi-cell MMSE-based precoding method that mitigates this problem. In addition to being linear, this precoding method has a simple closed-form expression that results from an intuitive optimization problem formulation. Numerical results show significant performance gains compared to certain popular single-cell precoding methods.

1.3 Organization

The rest of this dissertation has three chapters on network architecture, three chapters on interference management, and a concluding chapter.

In Chapter 2, we develop a rate-distortion framework for network resource allocation. In Chapter 3, we present a queueing architecture for cooperative networks, and perform stability analysis. In Chapter 4, we provide distributed rate allocation algorithms for wireless networks, and establish throughput-optimality results. The proof of lemmas in Chapter 3 are given in Appendix A and in Chapter 4 are given in Appendix B.

In Chapter 5, we present the genie-MAC outer bound for Gaussian interference networks. In Chapter 6, we present algorithms to perform robust weighted-sum rate maximization in MIMO interference networks. In Chapter 7, we analyze the pilot contamination in multi-cell TDD systems, and develop a precoding method that mitigate pilot contamination. The proofs of theorems

in Chapter 7 are given in Appendix C.

Finally, we conclude with Chapter 8.

1.4 Notation

$|\cdot|$ denotes the determinant of a square matrix, the cardinality of a set, and absolute value for scalars. $\mathbf{1}_{\{E\}}$ and $\mathbf{I}(E)$ denote the indicator function of event E. $(a)^+$ denotes $\max(a, 0)$. $\mathbb{E}[\cdot]$ and $\text{var}\{\cdot\}$ stand for expectation and variance operations, respectively. $\mathbb{R}_+$, $\mathbb{Z}_+$ and $\mathbb{Z}_{++}$ denote non-negative reals, non-negative integers and strictly positive integers, respectively.

Vectors and matrices are usually denoted by bold letters. For vectors, equality and inequality operators are defined component-wise. $\mathbf{a}\cdot\mathbf{b}$ denotes the dot product of $\mathbf{a}$ and $\mathbf{b}$, $\|\mathbf{a}\|_p$ denotes the L_p-norm, $\|\mathbf{a}\|_0$ denotes the number of non-zero elements. $\|\cdot\|$ denotes the two-norm. $\mathbf{0}$ denotes all-zeros vector, $\mathbf{1}$ denotes all-ones vector and $\mathbf{e}_i \in \{0, 1\}^n$ denotes the unit vector along i-th dimension.

$(\cdot)^T$ denotes the transpose and $(\cdot)^\dagger$ denotes the Hermitian transpose, $\text{tr}\{\cdot\}$ denotes the trace operation, $(\cdot)^{-1}$ denotes the inverse operation. $\|\cdot\|_2$ denotes the two-norm, and $\|\cdot\|_F$ denotes the Frobenius norm. $\mathbf{I}$ denotes the identity matrix. $\text{diag}\{\mathbf{a}\}$ denotes a diagonal matrix with diagonal entries equal to the components of $\mathbf{a}$. $(\mathbf{A})^{\text{upper}}$ denotes its upper triangular portion. $\mathbf{A} \succ 0$ denotes a symmetric positive-definite matrix.

Chapter 2

Network Resource Allocation: A Rate-Distortion Perspective

2.1 Introduction

Traffic forecasts predict mobile video to be the biggest portion of the world's mobile data in the near future [2]. Video is different from data and voice as video streaming quality can be varied with impact on user experience. Regardless, the vast majority of existing network architectures and protocols are designed for data. Even the majority of existing research on these topics assumes a packetized system, and then optimizes network performance [110].

An important component that is absent from such a framework is (lossy) compression. Compression is typically understood as an application-layer operation and thus separated from the network protocol stack optimization. However, the extent and nature of the compression employed critically impacts user experience, especially for video streaming. Assuming the sources are already quantized/compressed leads to a formulation that presents only a partial picture on the quality of service observed by the users in the system. For instance, lightly-compressed video may require rates much higher than those that can be allocated while ensuring stable network operation, while

heavily compressed video, although easy to deliver, reduces the quality of the end-user's experience. Thus, the distortion experienced by each user must be optimized to provide the best user experience (See [23, 27, 73] and references therein).

Integrating rate-distortion theory [10] into resource allocation is an important step from multiple perspectives, including multimedia applications, and thus bring elements of the application-layer into network optimization. There has been some prior work in this direction - for certain settings and alternate formulations, network operation optimization and rate-distortion theory have already been brought together. Rate-distortion optimized video streaming has been studied in the context of multimedia delivery, where the overall distortion incurred in the streaming process is dynamically minimized given changing network resources [27]. Similarly, optimal multiple description coding has also been studied from the networking perspective [79]. While each of these results have brought rate-distortion together with network constraints, a systematic analysis for compression is desirable, which is the main theme of this chapter.

2.1.1 Our Approach & Contributions

We build a new framework for network resource allocation by applying rate-distortion theory. Traditional resource allocation can be viewed as a special case where the distortion (and thus compression algorithm) is fixed at a value independent of network state and overall user experience. Distributed

compression problems have been studied and partially solved for special cases (such as Gaussian and/or binary sources) in particular settings. Typically, there is *no* provably optimal separation between source and channel coding in networks. However, for the special case of independent sources being transmitted through the network, it is known that separate source and channel coding is optimal [121]. Hence, we focus on independent (uncompressed) sources in the network that must be compressed and subsequently transmitted through the network. This applies to many scenarios including video streaming. Furthermore, focusing on source-channel separation allows us to develop a rate-distortion framework that scales with the network size, and hence, applicable to large real-world networks.

For networks with mutually independent (but possibly temporally correlated) sources, we consider the two quantities - (i) the *source-entropy* and (ii) its *distortion-offset* that are sufficient in representing compression. We formally define these two quantities in Section 2.2. Using these, we develop a framework for resource allocation. Few of the important implications of our framework are:

1. With lossy compression, the traditional notion of flow conservation does not hold. This has far-reaching consequences in network protocol design.

2. Our formulation based on *source-entropy* and *distortion-offset* has only linear constraints in addition to capacity constraints. Hence, if we focus on concave utility functions, existing convex optimization techniques can

be applied, especially the distributed algorithms developed in [74].

Based on the rate-distortion framework, we present the following results:

1. We show optimal decomposition of network resource allocation into three layers: (*i*) an application-layer with compression control, and (*ii*) a transport-layer with congestion control, and (*iii*) a network-layer with scheduling.

2. For a compression problem with binary sources and proportional-fair like utility functions, we derive the optimal policy. The optimal policy requires varying distortion based on link-rate, and hence, clearly shows the sub-optimality of decoupling compression problem from network optimization.

3. We optimally solve specific resource allocation problems involving sending binary and Gaussian uncompressed sources over multiple access channels.

2.1.2 Related Work

Incorporating compression into network optimization (or resource allocation) framework brings together different disciplines. The first of these is the domain of distributed lossy compression [10], a growing field of research. Distributed compression problems have been studied and partially solved for

special cases (such as Gaussian and/or binary sources) in particular settings. These include the multiple description problem [48], the CEO problem [95] and the two-terminal source coding problem [128]. These compression problems are formulated in an information-theoretic rate-distortion sense, where one or many sources must be compressed at minimal rates given distortion constraints. The resulting achievable *rate region* can be found for most multi-source multi-destination settings, and for a limited class of settings, shown to be optimal. Even though we focus on the setting where source-channel separation holds, the developed framework can be applied to general settings. Such a framework for capacity-constrained networks with correlated sources is studied in [32].

Over the years, we have gained a rich understanding of cross-layer optimization. The optimization problem formulation developed in [74] forms the foundation for our understanding of TCP (and rate control in general) as a solution to this optimization. Subsequently, multiple other network protocols have been formulated (and sometimes reverse-engineered) in terms of network optimization problems. The backpressure algorithm introduced in the context of stable operation of networks in [119] can be viewed as a dynamic solution to a similar optimization problem formulation. Indeed, rate control together with network stability can be formulated as an optimization problem [44]. It is known that a natural separation exists between the rate control mechanism and the network stability mechanism, and each of these problems can be individually solved and the solutions combined for optimal operation of networks.

Distributed solutions for rate control based on primal-dual methods can be found in [110].

The traditional resource allocation framework has been extended considerably to include other network features and characteristics. A significant fraction of this work is in incorporating the physical-layer aspects into the framework [93, 94, 108, 131]. Typically, signal to noise ratio (SNR) or signal to interference and noise ratio (SINR) based models have been used for this purpose [25]. An equal effort has been devoted to incorporating higher layer aspects into the problem structure, such as hierarchical network topologies [135], delay tolerant networks and so on. Cooperative networking strategies have also been studied in this context [106, 130]. Finally, the framework and the resulting optimization decomposition has been used to restructure the protocol stack and thus optimize overall system performance [96].

2.1.3 Organization

In Section 2.2, we present the rate-distortion framework for network resource allocation. In Section 2.3, we derive an optimal decomposition of resource allocation into layers. In Section 2.4, we study a compression control problem for binary sources. Finally, in Section 2.5, we apply the rate-distortion framework to Gaussian MACs.

2.2 Rate-Distortion Framework

We consider a single-hop[1] network with N independent sources, labeled $i = 1, 2, \ldots, N$. The i-th (possibly continuous-valued) source X_i has an *uncompressed-rate* of s_i symbols/sec. This source is compressed at a *distortion* of D_i (per symbol, averaged across time) to a *rate* of c_i bits/sec. In other words, a lossy-compression code exists that maps vectors comprised of source symbols to binary vectors such that recovery is possible to within a distortion of D_i per symbol. Mathematically, a rate-distortion code (operating over blocks of symbols of size n, with n large enough) of rate $c_i + \epsilon$ bits/sec exists for source X_i such that reconstruction to within a distortion D_i is possible such that $\epsilon \to 0$ as $n \to \infty$.

This compressed source is transmitted over a link with *link-rate* of r_i bits/sec. The corresponding vectors are denoted by $\mathbf{s}$, $\mathbf{D}$, $\mathbf{c}$ and $\mathbf{r}$, respectively. These link rates are coupled in a wireless network, and this, for a single-hop network, is captured by the N-dimensional information-theoretic rate region denoted by $\mathcal{C}$.[2] The parameters introduced so far are associated with different functionalities in a network: (i) s_i and D_i are associated with (lossy) source coding, (ii) c_i is associated with congestion (or rate) control, and (iii) r_i is associated with rate allocation (or scheduling).

The source coding, rate control and scheduling problems are closely tied

[1] This framework can be extended to multi-hop networks.

[2] This rate region may be the capacity region if the network's capacity region is known, or the best known rate region if unknown.

to each other. As a result, the parameters associated with these problems must be jointly optimized. Therefore, we desire a framework that simultaneously captures all these problems. However, the traditional framework does not include the source coding component. It is based on the following optimization problem:

$$\max_{\mathbf{r}} \quad \sum_{i=1}^{N} U_i(c_i) \tag{2.1}$$

subject to

$$c_i \leq r_i, \forall i, \tag{2.2}$$

$$\mathbf{r} \in \mathcal{C}. \tag{2.3}$$

In this framework, $U_i(c_i)$ in (2.1) is the (concave) utility function associated with the (compressed) rate c_i of i-th source and (2.2)-(2.3) are capacity constraints. This framework can be decomposed into two layers: a transport-layer performing rate control, and a network-layer performing scheduling [96].

To incorporate the source coding parameters, it is natural to utilize rate-distortion functions of sources studied in information theory [30]. For explaining this, we consider two source types:

1. **Binary sources with Hamming distortion:** Consider independent Bernoulli(p_i) binary sources that are mutually independent arriving at rates of s_i symbols per second. The rate-distortion function for this source is known to be

$$R(s_i, D_i) = s_i \left(H(p_i) - H(D_i)\right), \tag{2.4}$$

where $H(\cdot)$ is the binary entropy function given by

$$H(q) = -q \log_2 q - (1-q) \log_2 (1-q).$$

Now, motivated from (2.4), we define two variables to represent this source: (i) *source-entropy*

$$\alpha_i = s_i H(p_i) \tag{2.5}$$

in bits/sec, where s_i is the uncompressed-rate in symbols/sec and $0 < p_i < 1$ is the given Bernoulli parameter of i-th source, and (ii) (negative) *distortion-offset*

$$\beta_i = -s_i H(D_i) \tag{2.6}$$

in bits/sec, where D_i is the Hamming distortion per symbol.

2. **Gaussian sources with squared-error distortion:** Consider zero-mean independent Gaussian sources with variances σ_i^2 arriving at a rate of s_i symbols per second. With squared-error distortion, the rate-distortion function is known to be

$$R(s_i, D_i) = \frac{s_i}{2} \log_2 \frac{\sigma_i^2}{D_i}. \tag{2.7}$$

Now, differential *source-entropy* α_i and *distortion-offset* β_i are defined as follows:

$$\alpha_i = \frac{s_i}{2} \log_2 2\pi e \sigma_i^2,$$

where $\sigma_i^2 > 0$ is the given variance parameter of the i-th source, and

$$\beta_i = -\frac{s_i}{2} \log_2 2\pi e D_i,$$

where D_i is the squared-error distortion per symbol. Note that these two variables can take both positive and negative values.

Now, *source-entropy* and *distortion-offset* can be identified as two parts of the rate-distortion function for multiple types of sources, both i.i.d. and correlated (for example, see Shannon's rate-distortion lower bound [30]). This includes both binary and Gaussian sources as special cases. Denoting source-entropy and distortion-offset as α_i and β_i respectively, we have a tradeoff between the two of the form given by:

$$\alpha_i + \beta_i \leq c_i, \forall i. \tag{2.8}$$

This simply states that the compressed rate should be higher than the fundamental limit given by the rate-distortion function. Since distortion-offset terms appear in the constraints, it shows that flow conservation assumed in data networks does not hold for sources such as video. This motivates re-design of network protocol components that assume packets to be immutable.

Now, a user's happiness (or user experience) can be thought of as a function of the source-entropy and distortion-offset.[3] Therefore, a natural framework for network resource allocation is to maximize the sum of the user

[3]Since source-entropy and distortion-offset together is a one-to-one map from source-rate and distortion, there is no loss of generality.

experience subject to all network constraints. For deriving suitable layering architecture in the next section, we consider a slightly different looking but equivalent[4] framework for resource allocation:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{c},\mathbf{r}} \quad \sum_{i=1}^{N} V_i(\alpha_i, \beta_i) + U_i(c_i) \tag{2.9}$$

subject to

$$\alpha_i + \beta_i \leq c_i, \forall i, \tag{2.10}$$

$$a_i \alpha_i \geq 0, \forall i, \tag{2.11}$$

$$b_i \beta_i \leq 0, \forall i, \tag{2.12}$$

$$\alpha_i + \beta_i \geq 0, \forall i, \tag{2.13}$$

$$c_i \leq r_i, \forall i, \tag{2.14}$$

$$\mathbf{r} \in \mathcal{C}, \tag{2.15}$$

where $a_i, b_i \in \{0, 1\}$ are constants that are source-dependent, (2.10)-(2.13) are rate-distortion conditions and (2.14)-(2.15) are capacity constraints.

The rate-distortion framework in (2.9) has two main advantages.

- It presents a notion of optimal network optimization while dealing with uncompressed sources.
- The constraints in (2.10)-(2.14) are linear, and $\mathcal{C}$ in (2.15) is a convex set (with time sharing). Hence, with concave utility functions, we have a convex framework.

[4] $U_i(c_i)$ can be absorbed into $V_i(\alpha_i, \beta_i)$

For a convex framework, we consider utility functions with following two properties:

Definition 2.1 (Concave Utility)**.** *$V_i(\alpha_i, \beta_i)$ is jointly concave in α_i and β_i. $U_i(c_i)$ is concave in c_i.*

Definition 2.2 (Monotone Utility)**.** *Given a particular value of variable β_i (α_i), $V_i(\alpha_i, \beta_i)$ is monotone increasing in the other variable. $U_i(c_i)$ is monotone increasing in c_i.*

2.3 Decomposition into Multiple Layers

In this section, we show that the framework in (2.9) can be decomposed into three layers: (*i*) "application" layer with compression control, (*ii*) "transport" layer with congestion control, and (*iii*) "network" layer with (centralized) scheduling. As evident from the names, each of these layers has direct correspondence with a layer in the standard network protocol stack.

We proceed by introducing two sets of dual variables. We introduce non-negative dual variables $\mu_i, \forall i$ (vector denoted by $\boldsymbol{\mu}$) corresponding to constraints in (2.10), and non-negative dual variables $\lambda_i, \forall i$ (vector denoted by $\boldsymbol{\lambda}$) corresponding to constraints in (2.14). With these dual variable, we obtain the following Lagrangian:

$$\mathcal{L} = \sum_{i=1}^{N} V_i(\alpha_i, \beta_i) + U_i(c_i) - \sum_{i=1}^{N} \mu_i(\alpha_i + \beta_i - c_i) - \sum_{i=1}^{N} \lambda_i(c_i - r_i). \quad (2.16)$$

Now, the dual objective $g(\boldsymbol{\mu}, \boldsymbol{\lambda})$ is defined as

$$\begin{aligned} g(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{c},\mathbf{r}} \; & \sum_{i=1}^{N} V_i(\alpha_i, \beta_i) - \mu_i(\alpha_i + \beta_i) \\ & + \sum_{i=1}^{N} U_i(c_i) - (\lambda_i - \mu_i)c_i \\ & + \sum_{i=1}^{N} \lambda_i r_i \end{aligned} \tag{2.17}$$

subject to (2.11)-(2.13) and (2.15). From Langrange duality, it is well-known that $g(\boldsymbol{\mu}, \boldsymbol{\lambda})$ gives an upper bound on the primal problem in (2.9) for feasible primal and dual variables. This leads to the dual problem to obtain an upper bound on the primal problem, given by

$$\begin{aligned} \min_{\boldsymbol{\mu},\boldsymbol{\lambda}} \quad & g(\boldsymbol{\mu}, \boldsymbol{\lambda}) \\ \text{s.t.} \quad & \lambda_i \geq 0, \mu_i \geq 0, \forall i. \end{aligned} \tag{2.18}$$

For concave utility functions, under mild conditions [16], it follows that this dual problem is tight, i.e., the optimal value of (2.18) is equal to the optimal value of (2.9).

Now, it is fairly straightforward to see that the Lagrangian formulation in (2.17) decomposes into the following optimization problems:

1. **Distributed Compression Control:** For all i, given μ_i,

$$\begin{aligned} \max_{\alpha_i,\beta_i} \quad & V_i(\alpha_i, \beta_i) - \mu_i(\alpha_i + \beta_i) \\ \text{s.t.} \quad & a_i\alpha_i \geq 0, b_i\beta_i \leq 0, \alpha_i + \beta_i \geq 0. \end{aligned} \tag{2.19}$$

2. **Distributed Congestion Control:** For all i, given μ_i and λ_i,

$$\max_{c_i} \quad U_i(c_i) - (\lambda_i - \mu_i)c_i. \tag{2.20}$$

3. **Centralized MaxWeight Scheduling:** Given all λ_i,

$$\begin{aligned} \max_{\mathbf{r}} \quad & \sum_{i=1}^{N} \lambda_i r_i \\ \text{s.t.} \quad & \mathbf{r} \in \mathcal{C}. \end{aligned} \tag{2.21}$$

In contrast to traditional network optimization, the distributed compression problem in (2.19) is explicitly included in our decomposition. This problem jointly chooses source-entropy and distortion-offset based on the utility function. The congestion control in (2.20) and the centralized scheduling in (2.21) match with those known in existing literature [44, 110]. Note that, in general, all three problems in (2.19), (2.20) and (2.21) are coupled through dual variables $\boldsymbol{\mu}, \boldsymbol{\lambda}$. In many cases, it is possible to use gradient methods to solve for the dual variables [96]. Hence, we do not delve into a discussion of such methods to solve these problems. Instead, we focus on two problems to obtain further insights in combining compression control with resource allocation - first, we study a compression control problem for binary sources, and then, we apply our framework to Gaussian multiple access channels (MACs).

2.4 Compression Control for Binary Sources

Let us consider the lossy compression problem that determines source-entropy and distortion-offset given a compressed-rate. We study this problem

to understand the tradeoff involved in choosing higher source-rate (with higher distortion) versus lower source-rate (with lower distortion). With utility functions that are strictly increasing, it follows that optimal parameters satisfy $\alpha_i + \beta_i \leq c_i$ with equality. Under this setting, the compression control at every source is: for given a_i, b_i and c_i

$$\max_{\alpha_i} \quad V(\alpha_i, c_i - \alpha_i) \tag{2.22}$$

subject to

$$a_i \alpha_i \geq 0,$$
$$b_i(c_i - \alpha_i) \leq 0.$$

In order to obtain explicit solutions to the optimization problem in (2.22) , we solve it in the context of a binary source with Hamming distortion. For a binary source, we have $a = 1$ and $b = 1$. Consider the utility function[5]:

$$V(\alpha_i, \beta_i) = \log_e \alpha_i + K_i \beta_i, \tag{2.23}$$

for some constant $K_i > 0$. Note that this utility function is an extension of the proportional-fair utility function with linear penalty for distortion-offset. Therefore, (2.22) simplifies to

$$\begin{aligned} \max_{\alpha_i} \quad & \log_e \alpha_i + K_i(c_i - \alpha_i) \\ \text{s.t.} \quad & \alpha_i \geq c_i. \end{aligned} \tag{2.24}$$

[5]This is just an example, and the choice of utility functions that is appropriate in practice is a subject for further study.

The unconstrained problem in (2.24) is maximized at $\alpha_i = 1/K_i$. Therefore, for the constrained problem in (2.24), we have

$$\alpha_i^* = \begin{cases} 1/K_i, \text{ if } 1/K_i \geq c_i \\ c_i, \text{ otherwise.} \end{cases} \tag{2.25}$$

This simple rate-distortion-control policy can be implemented as long as the application layer is aware of the compressed-rate c_i.

The expression in (2.25) provides a simple rule to decide whether to transmit at zero-distortion, i.e., with source-entropy $\alpha_i = c_i$ and distortion-offset $\beta_i = 0$, or transmit with distortion, i.e., source-entropy $\alpha_i = 1/K_i$ and distortion-offset $\beta = c_i - 1/K_i$. When $1/K_i \geq c_i$, substituting $\alpha_i = 1/K_i$ and $\beta_i = c_i - 1/K_i$ in (2.5) and (2.6), respectively, we get the following: uncompressed-rate s in symbols/sec is given by

$$s_i = \frac{1}{K_i H(p)},$$

and Hamming distortion D_i is given by the expression

$$\frac{H(D_i)}{H(p)} = 1 - c_i K_i.$$

Recall that p is the Bernoulli parameter associated with source and $H(\cdot)$ is the binary entropy function. Thus, source-entropy and distortion-offset can be translated to the source coding parameters source-rate and distortion.

This compression rule is depicted in Figure 2.1. In simple words, this rule states that source coding with distortion has to be performed at low compressed-rates and source coding without distortion has to be performed

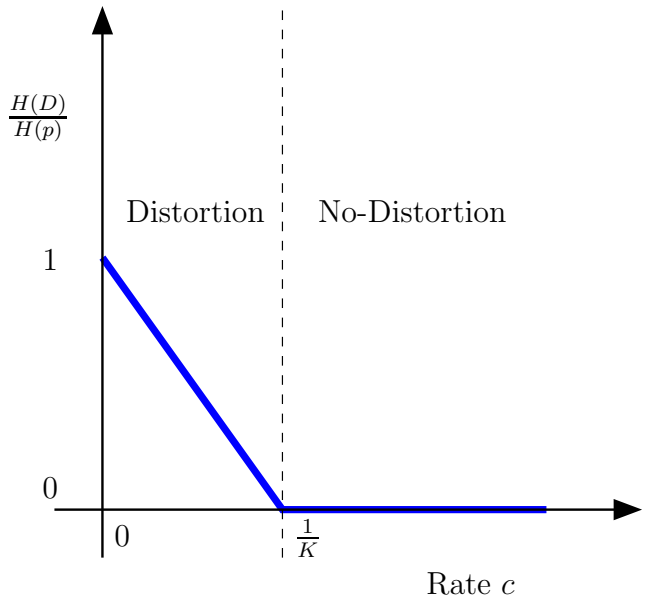

Figure 2.1: Compression control for binary sources; Region to the left of dashed line represents lossy source-coding while region to the right represents lossless source-coding

at high compressed-rates. Furthermore, the amount of distortion introduced by the compression algorithm is piecewise linear. This shows that the traditional approach of decoupling compression control from network optimization is suboptimal. In majority of existing video streaming systems, compression control is performed using ad hoc algorithms. The result in this section show that simple and optimal compression control can be developed.

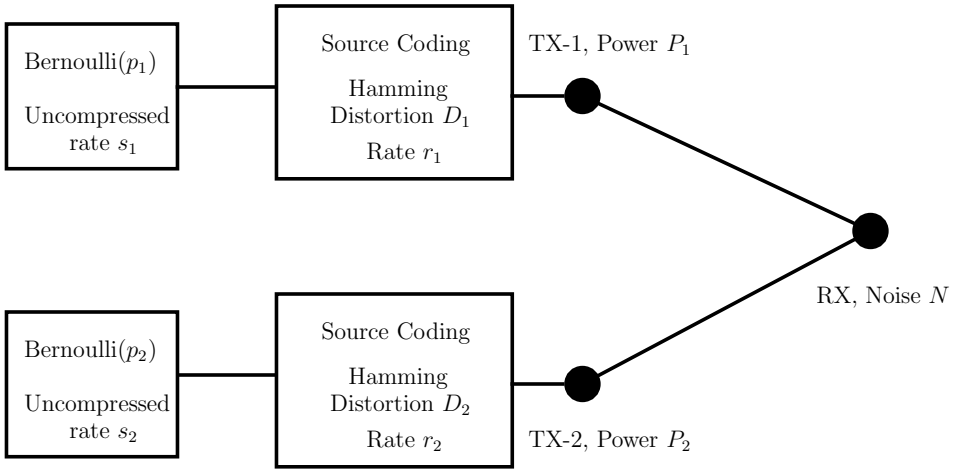

Figure 2.2: Multiple access channel with binary sources

2.5 Rate-Distortion Framework applied to Multiple Access Channels

Our next goal is to understand the interplay between compression and communication - specifically, the way channel capacity and resulting distortion impact one another. We choose Gaussian multiple access channel (MAC) for our analysis here as it represents the simplest multi-terminal system model, and the capacity region for a MAC is well known [30]. Further, we consider simple utility functions below that are only dependent on the distortion suffered in the compression process. These simplifications help us focus on our goal.

2.5.1 MAC with binary sources

Consider two i.i.d. Bernoulli(p_i) binary sources that are mutually independent (across sources) arriving at rates of s_i symbols per second. For a

binary source with Hamming distortion, the rate-distortion function is given by (2.4). The uncompressed-rates s_i are positive constants that are fixed by nature and assumed to be known. After compression, these two sources are to be communicated over a Gaussian multiple access channel as shown in Figure 2.2.

Now, the network resource allocation problem for this example can be expressed as

$$\max_{\mathbf{D}} \quad \sum_{i=1}^{2} V_i(D_i) \tag{2.26}$$

subject to

$$\begin{aligned} & s_i \left(H(p_i) - H(D_i)\right) \leq C(P_i), \forall i, \\ & \sum_{i=1}^{2} s_i \left(H(p_i) - H(D_i)\right) \leq C(P_1 + P_2), \\ & D_i \geq 0, D_i \leq 1, \forall i. \end{aligned}$$

Here, we have used the capacity region of the Gaussian MAC channel. $C(\cdot)$ corresponds to Shannon's capacity formula given by

$$C(P) = \frac{1}{2} \log_2 \left(1 + \frac{P}{N}\right).$$

Note that, if the utility function in (2.26) is concave in distortion, the optimization problem in (2.26) is in convex form.[6] This follows from the fact that entropy is concave.

[6]This is not in the convex form explained in Section 2.2.

Next, for deriving further insights into this problem, we consider the case where utility $V_i(D_i)$ in (2.26) is a linear function of $H(D_i)$, i.e.,

$$V_i(D_i) = -\delta_i H(D_i)$$

for some constant $\delta_i > 0$. With change of variables $x_i = s_i H(D_i)$, from (2.26), we obtain an equivalent linear program (LP) (with sign of optimal value reversed) given by

$$\min_{x_1,x_2} \quad \frac{\delta_1}{s_1} x_1 + \frac{\delta_2}{s_2} x_2 \tag{2.27}$$

subject to

$$\begin{aligned} & x_i \geq s_i H(p_i) - C(P_i), \forall i, \\ & x_1 + x_2 \geq s_1 H(p_1) + s_2 H(p_2) - C(P_1 + P_2), \\ & x_i \geq 0, x_i \leq s_i, \forall i. \end{aligned}$$

From properties of LP, it follows that at least one optimal solution exists that is a corner point of the feasible set, which is the convex polytope characterized by the constraints of the problem in (2.27). More intuitively, we can obtain the optimal corner points for different cases based on where the source entropy vector $\mathbf{H} = (s_1 H(p_1), s_2 H(p_2))$ lies with respect to the MAC capacity region $\mathcal{C}$:

1. Case-A ($\mathbf{H} \in \mathcal{C}$): The optimal corner point is $D_1^* = 0$, $D_2^* = 0$, i.e., perform lossless source coding.

2. Case-B ($\mathbf{H} \notin \mathcal{C}$): It follows from the MAC capacity region (and utility function) that there are only two corner points of interest. These are the corner points on the sum-capacity boundary. The exact corner points and the condition for choosing between these corner points are as follows: If $\delta_1/s_1 \geq \delta_2/s_2$, then

$$\begin{aligned} s_1 H(D_1^*) &= [s_1 H(p_1) - C(P_1)]^+ , \\ s_2 H(D_2^*) &= [s_1 H(p_1) - (C(P_1 + P_2) - C(P_1))]^+ , \end{aligned}$$

otherwise,

$$\begin{aligned} s_1 H(D_1^*) &= [s_1 H(p_1) - (C(P_1 + P_2) - C(P_2))]^+ , \\ s_2 H(D_2^*) &= [s_1 H(p_1) - C(P_2)]^+ . \end{aligned}$$

Here, $[x]^+$ denotes the positive part of x given by $\max\{0, x\}$.

Thus, we have explicitly solved the resource allocation problem for this illustrative example. We depict this solution in Figure 2.3. This figure captures the intuitive distortion-control policy: compute weights and choose the corner point for operation corresponding to the largest weight.

2.5.2 MAC with Gaussian sources

Next, we consider independent Gaussian sources with squared-error distortion. Using this example, we show that optimal distortion-control does not necessarily result in corner points corresponding to the capacity region, even for certain natural utility function. While using a decomposition approach,

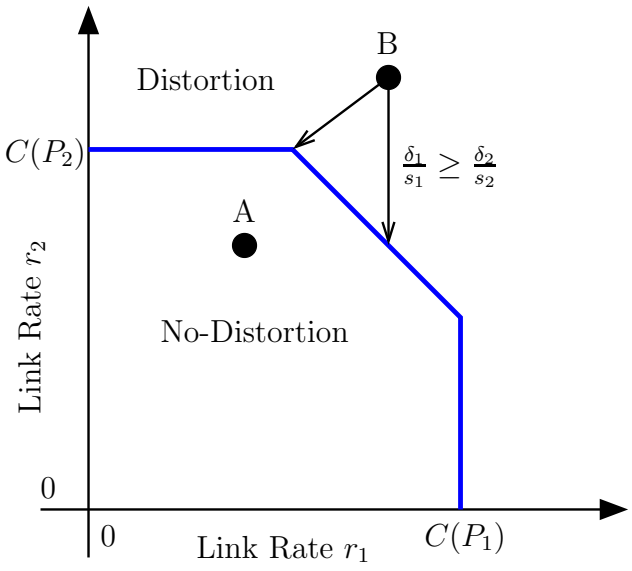

Figure 2.3: Optimal resource allocation for multiple access channel with binary sources; Point-A corresponds to Case-A (no-distortion), and Point-B corresponds to Case-B (distortion)

the max-weight scheduling component usually chooses one of the corner points. Therefore, using this example, we show that the decomposition approach leads to strictly sub-optimal solution. Since the objective considered does not result in a convex formulation, this is not surprising. However, it is interesting that some of these non-standard problems can be solved in closed-form.

Consider two i.i.d. Gaussian sources with variance σ_i^2 arriving at a rate of s_i symbols per second. These sources are to be communicated over a Gaussian MAC channel. Then, the optimal resource allocation problem is:

$$\max_{\mathbf{D}} \quad \sum_{i=1}^{2} V_i(D_i) \tag{2.28}$$

subject to

$$\begin{aligned}
&\frac{s_i}{2}\log\frac{\sigma_i^2}{D_i} \le C(P_i), \forall i,\\
&\sum_{i=1}^{2}\frac{s_i}{2}\log\frac{\sigma_i^2}{D_i} \le C(P_1+P_2),\\
&D_i \ge 0, \forall i.
\end{aligned}$$

Now, we consider linear utility function in distortion given by

$$V_i(D_i) = -\delta_i D_i. \tag{2.29}$$

It follows from (2.29) and (2.28) that the optimal max-weight scheduling lies on the sum-capacity facet. However, in general, it does not correspond to one of the corner points in this facet.

With change of variables to rates given by

$$r_i = \frac{s_i}{2}\log_2\frac{\sigma_i^2}{D_i}$$

and using the fact that for optimal rates, the constraint

$$r_1 + r_2 \le C(P_1+P_2),$$

is satisfied with equality, we obtain the following equivalent problem (optimal value scaled by a negative constant) for (2.28):

$$\begin{aligned}
\min_{r_1} \quad & \exp\left(-\frac{2r_1}{s_1}\right) + \gamma\exp\left(\frac{2r_1}{s_2}\right)\\
\text{s.t.} \quad & r_1 \le C(P_1),
\end{aligned} \tag{2.30}$$

where

$$\gamma = \frac{\delta_2 \sigma_2^2}{\delta_1 \sigma_1^2} \exp\left(-\frac{2C(P_1+P_2)}{s_2}\right).$$

Now, by differentiating the function in (2.30) w.r.t. r_1 and equating to zero, we get

$$-\frac{2}{s_1}\exp\left(-\frac{2\hat{r}_1}{s_1}\right) + \frac{2}{s_2}\gamma \exp\left(\frac{2\hat{r}_1}{s_2}\right) = 0,$$

which simplifies to

$$\hat{r}_1 = \frac{s_1}{s_1+s_2}C(P_1+P_2) + \frac{s_1 s_2}{2(s_1+s_2)}\log\left(\frac{\delta_1\sigma_1^2 s_2}{\delta_2\sigma_2^2 s_1}\right).$$

It is straightforward to check that the second derivate of the function in (2.30) w.r.t. r_1 is strictly positive at this point. For the constrained problem in (2.30), using elementary functional analysis, it turns out that the optimal solution is $r_1^* = \min\{\hat{r}_1, C(P_1)\}, r_2^* = C(P_1+P_2) - r_1^*$. For a symmetric case (i.e., all parameters associated with the two sources are equal), the above solutions leads to equal rates for both links, i.e., $r_1^* = r_2^* = C(P_1+P_2)/2$.

The above result suggests that, from a distortion-control perspective, a max-weight scheduling policy for choosing operating points on the capacity region is not always sufficient. However, if we restrict focus to concave utilities, a max-weight scheduling policy is sufficient.

Chapter 3

Queue-Architecture and Stability Analysis in Cooperative Relay Networks

3.1 Introduction

Cooperative relaying is traditionally seen as a physical-layer communication scheme involving multiple nodes in the network [92], with limited network-layer insights originating from such schemes. Indeed, the not-so-uncommon perception is: whatever be the physical-layer transmission/coding scheme, the network can abstract it into a *rate region* and then determine algorithms to stabilize queues, perform rate control and other tasks at the higher layers. From this perspective, it seems unimportant to learn about the intricacies in a multi-hop multi-user communication scheme.

There is a significant and growing body of work suggesting that such abstractions may not be accurate [36] and that physical-layer aspects must be included into the analysis. A large class of this work is based on signal-to-noise ratio (SNR) or signal-to-interference-and-noise ratio (SINR) models for the physical medium. While this is a worthwhile abstraction for physical-layer schemes that *treat interference as noise*, it is often overused and does not capture more involved physical-layer transmission schemes. From information

theory, it is well known that treating interference as noise represents a very limited class of transmission schemes, and a much larger class of schemes exist that achieve significantly higher throughput. Therefore, a framework that brings the information-theoretic coding scheme together with network-stability analysis is needed, to bridge the gap caused by the "unconsummated union" [38].

We explore building this bridge in the context of cooperative relay networks. For cooperative relay networks, multiple reasons exist for jointly studying queueing and cooperative communication schemes.

- First, the rate-maximizing physical-layer coding strategy automatically imposes scheduling restrictions on the relays/transmitters in the network. For coherent combination at the receivers to be at all possible, all nodes involved must transmit simultaneously in that block.

- Second, it is codebooks and functions of codebooks being received, stored and transmitted by nodes and not traditional data packets.

- Finally, the codebook chosen by the source(s) determines the rate of transmission, which may or may not be alterable at intermediate nodes (this is a key distinction between general information-theoretic coding theorems and say, packetized or linear network coded systems where rate can always be varied at every node). For example, if a relay were to use amplify-and-forward or compress-and-forward as its physical-layer strategies, it has no control over rate and has a real vector as its "packet".

3.1.1 Organization

In Section 3.2, we present a brief background on cooperative relay networks. In Section 3.3, we present our main results. In Section 3.4, we describe our system model in the context of heterogeneous cellular networks. In Section 3.5, we describe cooperative schemes for such networks in detail and present a queue-architecture that enables both efficient and optimal operation of the network. In Section 3.6, we present the main algorithm for operating such networks, and establish that this algorithm is throughput-optimal.

3.2 Background on Cooperative Relay Networks

Cooperative relay networks have been researched extensively since the "MIMO effect" was established. Until recently, it was considered hard if not impractical for nodes to coordinate transmissions to enable cooperative relaying. However, emerging heterogeneous cellular networks are increasingly moving in the direction of standardizing and evaluating schemes with node cooperation [1]. As cell sizes decrease, an increase in cell edges and interference requires node cooperation to increase throughput, and cooperative relaying is an important step in making this happen.

Figure 3.1 shows a basic configuration that incorporates cooperative relaying in heterogeneous cellular networks. To motivate this setting, we take the example of a macro-cellular network. Here, the source node S corresponds to the macro-cell base-station, the relay nodes R_1 and R_2 correspond to pico-cell base-stations and the destination nodes D_1, D_2 and D_3 correspond to

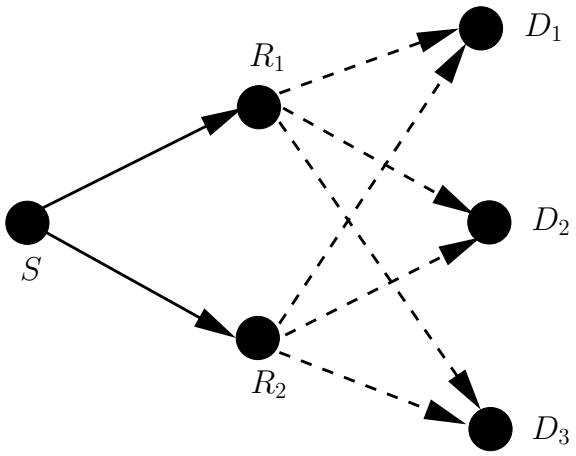

Figure 3.1: A two-hop cooperative network

mobiles. We focus on the downlink scenario where the source S has independent messages/bits for the mobiles. The relays' role is to help the source in transmitting these messages. Further, we assume a half-duplex cooperative constraint so that either the first-hop or the second-hop links can be activated at any given time, with no direct-links from the source to the destinations. A more general and detailed system model for such cooperative relay networks is provided in Section 3.4.

Even for simpler networks such as two relays and one destination and fixed channels, information-theoretic capacity is not yet known. However, there has been significant progress in developing cooperative communication schemes for such systems by using coherence and physical-layer coordination among nodes [77, 107]. There are multiple strategies studied in literature that enable this coordination, referred to as *forwarding* schemes [24, 31, 104]. One such scheme of interest is the so-called decode-and-forward scheme that re-

quires relays to decode messages. In contrast to traditional networks, the relays decode common messages that are then transmitted cooperatively. However, the relays still have decoded messages or packets as in traditional networks. In [132], the authors develop a throughput-optimal network algorithm that can handle common messages. In [133], the authors consider more general network configurations, but the applicability is still limited to decode-and-forward schemes with fixed channels. In essence, all of these apply only in packet-in packet-out networks. Complimentary to this is the work on optimal resource allocation for non-cooperative wireless networks [44, 80, 120] (and references therein).

For enabling generic cooperative schemes, we need to go beyond a packet-in packet-out framework. We desire that the relays use *any* information-theoretic cooperative coding strategy of its choice, be it amplify, compress, quantize or any alter-and-forward scheme. This couples cooperative coding and queueing into one joint problem, and the analyses in [44, 120, 132] and the vast literature on non-cooperative networks do not apply. Even the analyses in [132, 133] for decode-and-forward cooperative networks do not apply. This motivates the need for a new framework and stability analysis.

3.3 Our Approach & Main Results

Before proceeding to describe our results, a note to state the obvious: if the channel state is fixed and thus its capacity is precomputed, a simple static split scheme will ensure stable operation while maximizing the information-

theoretic rate (region) for the network. The challenge, of course, is when the fading state distribution and input arrival rates are unknown, and the fading state can only be observed causally. Consider a fading channel with block fading of T symbols each. When T is much smaller compared to the packet duration (or equivalently the channel-coding duration), queueing/buffering of packets at relays is not required as the first-hop and second-hop can be operated sequentially without reducing data-rates. When T is comparable to (or larger than) the packet duration, queueing of packets at relays can provide significant gains in terms of data-rates. Furthermore, when T is roughly the same as the packet duration, queueing at relays is inevitable as the source does not know the fading state of the second-hop while encoding the packet. We focus on the second scenario when T is larger than the packet/codeword duration. Given that the channel distribution is unknown and the fading state is only known causally, we ask the question: Is it possible to stabilize the network while operating it close to the boundary of its information-theoretic rate region?

The answer to the preceding question is *yes*, which we first proved for a simpler network with two relays and one destination in [71]. For cooperative schemes such as amplify/quantize-and-forward and partial-decode-and-forward, the relays receive and transmit real-valued *packets*. In [71], we introduce a new *state-based* queue-architecture for these real-valued packets, and develop a throughput-optimal network algorithm that does not require the knowledge of the fading distribution. Each *state* corresponds to a vec-

tor comprised of the *entire* channel-state of each link in the network. This approach, although analytically very helpful, suffers from a drawback - it requires that a virtual-queue be maintained for each channel-state at each node in the network. This leads to an explosion of queues, even for simple network configurations. Moreover, the approach in [71] is specific to a single destination setting. Hence, we develop a simpler queue-architecture to enable stable operation of cooperative relay networks and generalize it to any number of destinations.

The queue-architecture presented is primarily *encoding-based*. This architecture is motivated by the manner in which adaptive modulation and coding is currently implemented in practice. In a majority of systems, the source node implements a limited number of encoding schemes (encoding functions and rate-vectors). Each encoding scheme is designed so that it can be successfully employed for a particular subset of states. Even though encoding schemes belong to a finite (and usually small) set, the mapping functions at the relays and the decoding functions at the destinations are usually state-dependent. A queue-architecture that keeps virtual-queues at the relays for each state corresponding to the first-hop and each encoding scheme is sufficient. This considerably reduces the number of virtual-queues that must be maintained while still remaining a "sufficient statistic", i.e., these encoding-based queues are a sufficiently rich-enough representation for us to develop throughput optimal algorithms.

Using this new and somewhat intuitive queue-architecture, we develop

a network algorithm that has the following properties.

1. It does not require the knowledge of the fading distribution.
2. It does not require the knowledge of the arrival rates.
3. It keeps all the queues stable for any arrival rate-vector within the throughput region, i.e., it is throughput-optimal.

Note that limiting ourselves to a small set of possible encoding schemes and rates inherently reduces the network's information-theoretic rate region. This loss in rate region would be smaller with more fine-grained encoding schemes and resulting queue-architecture. The encoding-based queue-architecture itself does not introduce any sub-optimality.

To sum up, we introduce and study a new encoding-based queue-architecture, which is inspired by an adaptive coded modulation system analyzed and implemented at the physical-layer in systems today. There is rich literature that show the need for interaction between network-layer algorithms and adaptive coding/modulation in emerging wireless networks [44, 80, 120]. When it comes to cooperative relay networks, this coupling between network-layer algorithms and adaptive coding/modulation is even more important and intricate. Specifically, we show that new queuing architectures are required to develop throughput optimal algorithms that can achieve any information-theoretic rate region corresponding to its choice of encoding/decoding strategies while maintaining stability.

3.4 System Model

We consider discrete-time two-hop cooperative networks that include the network shown in Figure 3.1. We allow for arbitrary number of relays and destinations, i.e, the network consists of a source node denoted by S, N relay nodes denoted by $R_1, R_2, \ldots, R_N$, and K destination nodes denoted by $D_1, D_2, \ldots, D_K$. The source has independent messages for all the destinations. The relays assist in transmitting these messages to their respective destinations. Throughout this chapter, *first-hop* refers to the links from the source to the relays, and *second-hop* refers to the links from the relays to the destinations. At any given time, half-duplex and cooperative-communication constraints require that either the first-hop or the second-hop can be activated and not both. The presence of direct links from source to destinations will not invalidate the analysis presented, but would allow for more involved cooperative schemes. For simplicity, in our system model, we assume that direct links are absent and thus concentrate on equal-path length networks.

Remark 3.1. *If the direct links are weak, these can be ignored without significant impact on the throughput region. The cooperative schemes developed assuming no direct links will remain applicable. If the direct links cannot be ignored, highly sophisticated cooperative schemes such as block Markov coding are required. In addition to the coding complexity, these coding schemes require knowledge of the direct links. The analysis can be extended to this scenario by including direct links into the state (defined later), but the lower complexity of the queue-architecture is obtained by exploiting the absence of direct links.*

The channel model does not directly impact the queue-architecture, and thus the network algorithm and stability analysis presented. Consider any state dependent channel.[1] The joint state-distribution is unknown a-priori but the instantaneous realizations are known causally. A particular channel model of interest is a linear interaction model with additive white Gaussian noise (AWGN). In the context of an AWGN channel, an example of state is a multiplicative fading parameter. We focus on a framework with i.i.d. block-fading model with a block-length of T symbols. The channels remain constant for the duration of one block, and then change to a new (independent) realization from an underlying distribution from block to block. Let $t \in \mathbb{Z}_+$ denote the channel fading blocks, and let $\mathcal{F}$ denote the fading state-space,[2] which is assumed to be discrete. In block t, $\mathbf{f}_1[t] \in \mathcal{F}^N$ denotes the fading realization for the first-hop and $\mathbf{f}_2[t] \in \mathcal{F}^{NK}$ denotes the fading realization for the second-hop. The combined fading-state is denoted by $\mathbf{f}[t] = (\mathbf{f}_1[t], \mathbf{f}_2[t])$. The corresponding random vectors are denoted by $\mathbf{F}_1[t]$, $\mathbf{F}_2[t]$ and $\mathbf{F}[t]$. Note that $\mathbf{F}[t]$ is i.i.d. over time, but can be spatially correlated. Let the probability that $\mathbf{F}[t]$ takes value $\mathbf{f}$ be $\pi_{\mathbf{f}}$. This is the underlying probability distribution that is unknown to the central controller.

Next, we explain the time-scales in which network and channel parameters evolve in our system. The coherence time T is assumed to be comparable to the channel-coding length in symbols. For the ease of presentation, "packet"

[1] A state dependent channel is modeled by a conditional probability distribution of the output given the input and the state.

[2] In the context of fading AWGN, $\mathcal{F}$ is a finite set of positive real-valued SNRs.

(which is either the channel codeword or any real-vector representing the actual data packet) length is assumed to be equal to the coherence time T. It is straightforward to extend the analysis when the packet length is a sub-multiple of the coherence time T. Each packet is transmitted on the first-hop and the second-hop exactly once. These transmissions need not happen in consecutive time-blocks, i.e., these packets can be buffered at the relays. The coding performed at the source, the mappings performed at the relays, and the decoding at the destinations can be arbitrary, i.e., this includes any and all schemes that are information-theoretically capacity-optimal or, if capacity is unknown, then the best known coding scheme. Further, we assume that the instantaneous fading-state is causally known to the central controller. In other words, prior to transmission, the central controller is aware of the entire network channel state for that particular time-block.

At the source node S, there are K queues consisting of bits (or data) corresponding to the K destinations. We denote the queue at the source corresponding to k-th destination by Q_S^k with queue-length $Q_S^k[t]$ during block t. There is an exogenous i.i.d. arrival process $A^k[t]$ of data-bits into Q_S^k with mean rate $\lambda_k T$ bits/block and bounded variance. The vector of arrival rates λ_k is denoted by $\boldsymbol{\lambda}$. We need a systematic technique for buffering encoded packets at the relays so that efficient adaptive control can be performed. We refer to such a buffering technique as queue-architecture, which is the main topic of the next section.

3.5 Queue-Architecture & Throughput Region

The notion of a "packet" in cooperative schemes is different from traditional networks where a packet is decoded at all intermediate relays, and is usually meant for one destination. In our terminology, the term "packet" refers to the set of coded symbols transmitted/received in the network. Note that each of the relays receives a different noisy version of the transmitted vector (transmitted "packet"), which is subsequently mapped to a transmit vector ("packet") at each relay. Again, the destinations receive a noisy version of a linear combination of relays' transmit "packets". We refer to the physical-layer signalling vectors as *packets*[3] at each node in the network.

Consider a packet that is transmitted from the source to the K destinations. Let this packet be transmitted on the first-hop during block t_1, and be transmitted on the second-hop during block t_2. Then, $\mathbf{g} = (\mathbf{f}_1[t_1], \mathbf{f}_2[t_2])$ is said to be the *state*[4] seen by this packet. A packet transmitted by the source is received by all the destinations in two hops, but the amount of information each destination receives varies depending on the encoding rates. Given a state seen by the packet, the set of encoding rates that can be supported is known as the rate region for the given state. An extremely challenging problem even in the single destination setting is to find the set of all achievable rates, or the capacity region for the given state. Even though the capacity region is

[3]We choose to use this language as the entire network-layer analysis is based on understanding the dynamics of these transmit vectors as they traverse the system.

[4]Note that this notion of state is different from physical channel fading state, but is it of equal importance in our analysis.

unknown in most cases, there are many efficient cooperative communication schemes that have been developed. Therefore, our main aims are: (i) to develop a queue-architecture that can support existing (and future) cooperative schemes, and (ii) to develop a throughput-optimal network algorithm using this queue-architecture.

3.5.1 State-based Queue-Architecture

The queue-architecture we developed in [71] for single-destination setting keeps virtual-queues at relays for every state. Next, we describe this using amplify-and-forward coding scheme as it facilitates the understanding of the new queue-architecture introduced in Section 3.5.2.

Amplify-and-Forward Scheme: Consider the single destination setting with AWGN channels. Further, consider an average power constraint of P per block per node in the network, and additive Gaussian noise of unit variance at each receiver in the network. If the source transmits during fading state $\mathbf{f}$, then the received signals at the relays are

$$y_n = \sqrt{f_{1,n}} x_S + w_n, \text{ for } n = 1, \ldots, N, \tag{3.1}$$

where x_S denotes the symbol transmitted from the source, and y_n denotes the symbol received at the relay R_n. Instead, if the relays transmit during fading state $\mathbf{f}$, then the received signal at the destination is

$$y_D = \sum_{n=1}^{N} \sqrt{f_{2,n}} x_n + w_D, \tag{3.2}$$

where x_n denotes the symbol transmitted from the relay R_n, and y_D denotes the symbol received at the destination. Here, w_n and w_D are i.i.d. zero-mean additive Gaussian noise of unit variance at the relay R_n and the destination D.

In the static case without any link activation constraint, amplify-and-forward commonly refers to the relaying scheme at the relays that transmit (in every time slot) scaled versions of the received signals in the previous time slot. The scaling parameters at the relays are determined as a function of the signal-to-noise ratios (SNRs). These scaling parameters should also ensure that the average power constraints at the relays are satisfied. We look at an amplify-and-forward scheme (denoted by AF scheme) in which the relays can transmit any of the previously received signal vectors or choose not to transmit. We assume that received signal-vectors at the relays are transmitted to the destination only once.

We say that a symbol x_S is transmitted by the source to the destination over state $\mathbf{g} = (g_{1,1}, \ldots, g_{1,N}, g_{2,1}, \ldots, g_{2,N}) \in \mathcal{F}^{2N}$, if the source transmits during a fading state of the form $(g_{1,1}, \ldots, g_{1,N}, *, \ldots, *)$ and the relays transmit during a fading state $(*, \ldots, *, g_{2,1}, \ldots, g_{2,N})$, where $*$ can be any fading level. Consider a symbol x_S transmitted by the source to the destination over some state $\mathbf{g}$. Let the average power used at the source S be $P_S^{\mathbf{g}}$ and at the relay R_n be $P_n^{\mathbf{g}}$. These parameters are later optimized for the state $\mathbf{g}$. From (3.1),

(3.2), the received symbol at the destination is

$$y_D = \sum_{n=1}^{N} \left[\sqrt{\frac{g_{2,n} P_n^{\mathbf{g}}}{g_{1,n} P_S^{\mathbf{g}} + 1}} \left(\sqrt{g_{1,n}} x_S + w_n \right) \right] + w_D, \tag{3.3}$$

where x_S has zero mean and variance $P_S^{\mathbf{g}}$. From (3.3), it is straightforward to see that the maximum rate we can obtain is

$$r_{\mathbf{g}} = \max_{P_S^{\mathbf{g}}, P_1^{\mathbf{g}}, \ldots, P_N^{\mathbf{g}} \leq P} C \left(\frac{P_S^{\mathbf{g}} \left(\sum_{n=1}^{N} \sqrt{g_{1,n} c_1} \right)^2}{1 + \sum_{n=1}^{N} c_n} \right), \tag{3.4}$$

where $c_n = (g_{2,d} P_n^{\mathbf{g}})/(g_{1,n} P_S^{\mathbf{g}} + 1)$ and $C(x) = \frac{1}{2} \log_2(1 + x)$.

Remark 3.2. *The rate in (3.4) is equal to the maximum achievable rate using amplify-and-forward scheme in the static case with fixed channel state* $\mathbf{g}$ *and full-duplex operation [104]. The power optimization in (3.4) will result in utilizing maximum power at the source and one of the relays. In general, the other relays will result in using lower power than the maximum available power [104].*

We can obtain rates *strictly* greater than the average of rates over all fading states by buffering at relays. Buffering enables optimal combining of states between the source and the relays and the relays and the destination. We will demonstrate this using a simple example. Let $P = 1$, $\mathcal{F} = \{0, 1, 10\}$. Consider the fading distribution such that fading states $(0, 0, 0, 0)$, $(0, 0, 10, 10)$, $(1, 1, 0, 0)$ and $(1, 1, 10, 10)$ occur with probabilities γ^2, $\gamma\bar{\gamma}$, $\bar{\gamma}\gamma$, and $\bar{\gamma}^2$, respectively. Here, $\bar{\gamma} = 1 - \gamma$ and γ can be viewed as the probability of outage in this example. The rate corresponding to state $(1, 1, 10, 10)$ alone is non-zero,

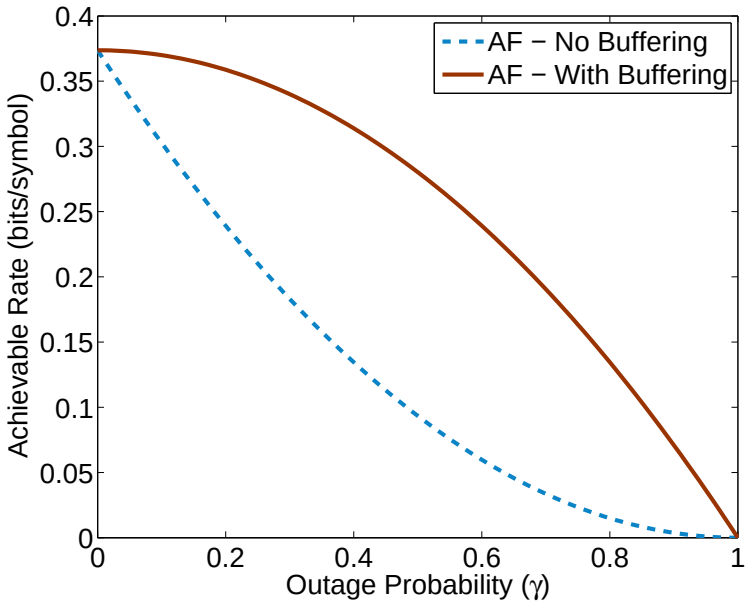

Figure 3.2: Throughput improvement via buffering at relays

which is $C(20/11)$. In this example, it is easy to observe that we can achieve $(0.5\bar{\gamma}^2+\gamma\bar{\gamma})C(20/11)$ bits/transmission with buffering whereas $0.5\bar{\gamma}^2C(20/11)$ bits/transmission is the average of rates over different fading states. These achievable rates are plotted in Figure 3.2.

State-based Queue-Architecture: For handling variable rate allocation, we maintain separate virtual queues at each relay based on the possible rates of the real-valued "packet". This is necessitated by the fact that encoding and decoding in amplify-and-forward relaying is an end-to-end process. Let $Q_n^{\mathbf{g}}[t]$ denote the virtual queue maintained for state $\mathbf{g}$, where $\mathbf{g} = (\mathbf{g}_1, \mathbf{g}_2) \in \mathcal{F}^4$. Note that when the system is in fading state $\mathbf{f}$, for any state $\mathbf{g}$ such that $\mathbf{g}_1 = \mathbf{f}_1$, a "packet" generated with rate $r_{\mathbf{g}}$ can be successfully transmitted

from the source to the relays; and for any state $\mathbf{g}$ such that $\mathbf{g}_2 = \mathbf{f}_2$, a "packet" generated with rate $r_{\mathbf{g}}$ can be successfully transmitted from the relays to the destination.

For the multiple destinations setting, suppose that each rate-region can be quantized such that the convex-hull of the set of quantized rate-vectors is "nearly" same as the rate-region itself. Further, let us assume that the rate corresponding to each destination have to be quantized to L levels. Now, a direct extension of the state-based queue-architecture would require virtual-queues at relays for each state and each quantized rate-vector, which results in $L^K|\mathcal{F}|^{K(N+1)}$ virtual-queues. This scales exponentially in the number of destinations K. Clearly, such a queue-architecture is not scalable, and might face implementation issues unless we take advantage of specific structure present in the system. This is the topic we explore next.

3.5.2 Encoding-based Queue-Architecture

In order to design a low-complexity queue-architecture, we exploit the fact that practical systems implement limited number of encoding schemes, as in the case of adaptive modulation and coding. For example, the source might choose to encode only two destinations at a time using superposition encoding. In this case, the total number of encoding schemes would be $K(K-1)L^2$. In another example, the source might choose to encode at limited boundary rate-vectors again with superposition encoding. Let $\mathcal{M}$ denote the set of encoding schemes, and $\mathbf{r}_m$ denote the rate-vector corresponding to each encoding scheme

$m \in \mathcal{M}$. Given that $|\mathcal{M}| \ll L^K |\mathcal{F}|^{KN}$, a queue-architecture needs to support these limited choices. While a queue-architecture can take advantage of this, it needs to allow for arbitrary mapping at the relays and decoding at the destinations. These are usually state-dependent, for example, an amplify-and-forward mapping is state-dependent.

Before describing our queue-architecture, we characterize the throughput region of the two-hop cooperative network, which is definted next.

Definition 3.1 (Throughput Region)**.** *A rate-vector is achievable if there exists a static split scheduling scheme that supports it, i.e., achievable with the knowledge of the fading distribution. The throughput region is defined as the closure of the set of all achievable rate-vectors.*

Define $\mathcal{I} = \{(m, \mathbf{g}) | m \in \mathcal{M}$ can be supported by state $\mathbf{g} \in \mathcal{F}^{(N+1)K}\}$, which represents whether an encoding scheme is supported by a state or not.[5] Now, let $\mathbf{f} = (\mathbf{f}_1, \mathbf{f}_2)$ be any fading-state where $\mathbf{f}_1$ is the fading-state of first-hop and $\mathbf{f}_2$ is the fading-state of second-hop. Similarly, let $\mathbf{g} = (\mathbf{g}_1, \mathbf{g}_2)$ by any state. We define $\hat{\mathcal{F}} = \mathcal{F}^{(N+1)K}$, $\mathcal{I}_1 = \{(\mathbf{f}, \mathbf{g}) | \mathbf{g}_1 = \mathbf{f}_1\}$, and $\mathcal{I}_2 = \{(\mathbf{f}, \mathbf{g}) | \mathbf{g}_2 = \mathbf{f}_2\}$. Now, any static split rule can be represented using the following two sets of parameters: $a_{\mathbf{f}}^{m,\mathbf{g}}$ - the fraction of time for which packets corresponding to encoding scheme m and state $\mathbf{g}$ is transmitted from the source to the relays when the system is in fading state $\mathbf{f}$, and $b_{\mathbf{f}}^{m,\mathbf{g}}$ - the fraction of time for which these

[5] We do not explicitly deal with packet error rate, as it is assumed that the achievable rate-vector is defined appropriately with required packet error rate.

packets are transmitted from the relays to the destinations. The throughput region of the network is characterized in the following lemma.

Lemma 3.1. *A rate-vector* $\hat{\mathbf{r}}$ *is in the throughput region denoted by* $\mathcal{T}$ *only if there exists* $a_{\mathbf{f}}^{m,\mathbf{g}} \geq 0$ *and* $b_{\mathbf{f}}^{m,\mathbf{g}} \geq 0$ *for all* $m \in \mathcal{M}$, $\mathbf{g} \in \hat{\mathcal{F}}$ *and* $\mathbf{f} \in \hat{\mathcal{F}}$ *such that*

$$\hat{\mathbf{r}} = \sum_{m,\mathbf{g},\mathbf{f}} \pi_{\mathbf{f}} a_{\mathbf{f}}^{m,\mathbf{g}} \mathbf{r}_m \mathbf{1}_{\{(\mathbf{f},\mathbf{g})\in\mathcal{I}_1\}} \mathbf{1}_{\{(m,\mathbf{g})\in\mathcal{I}\}}, \tag{3.5}$$

$$\sum_{\mathbf{f}\in\hat{\mathcal{F}}} \pi_{\mathbf{f}} a_{\mathbf{f}}^{m,\mathbf{g}} \mathbf{1}_{\{(\mathbf{f},\mathbf{g})\in\mathcal{I}_1\}} = \sum_{\mathbf{f}\in\hat{\mathcal{F}}} \pi_{\mathbf{f}} b_{\mathbf{f}}^{m,\mathbf{g}} \mathbf{1}_{\{(\mathbf{f},\mathbf{g})\in\mathcal{I}_2\}}, \forall (m,\mathbf{g}) \in \mathcal{I}, \tag{3.6}$$

$$\sum_{m,\mathbf{g}} a_{\mathbf{f}}^{m,\mathbf{g}} + b_{\mathbf{f}}^{m,\mathbf{g}} \leq 1, \forall \mathbf{f}. \tag{3.7}$$

Proof. Please see Appendix A.1. □

An immediate corollary of this lemma is the following.

Corollary 3.2. *The throughput region* $\mathcal{T}$ *is convex.*

Encoding-based Queue-architecture: At each relay (say n), we keep virtual queues corresponding to each encoding scheme m and each fading state for the first-hop $\mathbf{g}_1$ denoted $Q_n^{m,\mathbf{g}_1}$ with queue-length $Q_n^{m,\mathbf{g}_1}[t]$ during block t as shown in Figure 3.3. This queue consists of real-valued packets encoded at rate $\mathbf{r}_m$, i.e., these queue length units are not bits, but symbols. Since we keep virtual queues for every fading state corresponding to the first-hop, the mapping function performed at the relays can be a function of the state. Consider a set of packets chosen for transmission from the relays to the destinations. Now, the first-hop state information is known from the virtual

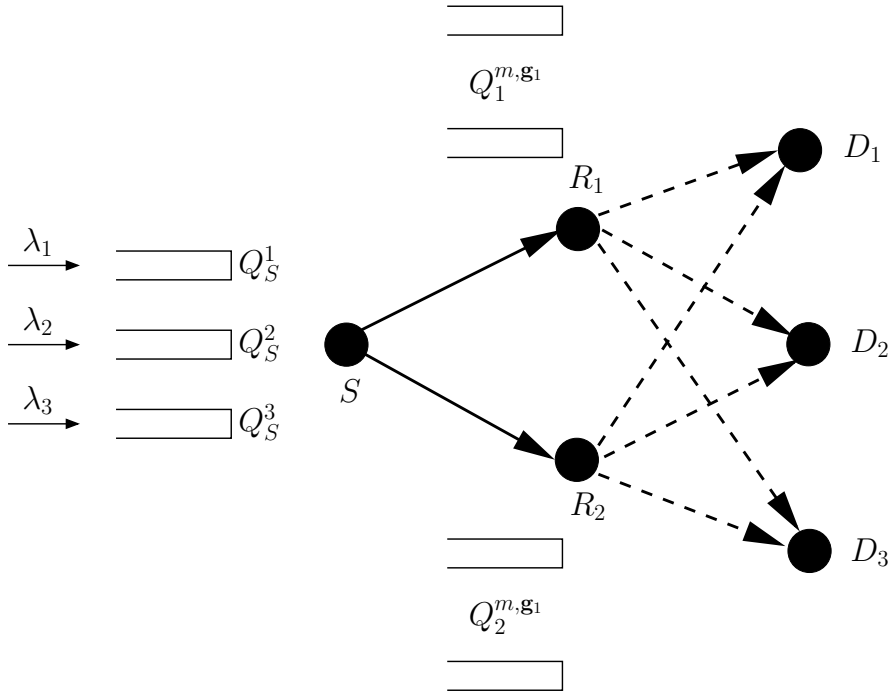

Figure 3.3: Encoding-based queue-architecture; λ_k is the arrival rate for the queue corresponding to destination D_k; m represents encoding scheme and $\mathbf{g}_1$ represents first-hop state

queues these belong to and the second-hop state is the current fading realization for the second-hop. Thus, the entire state is known while the relays perform this transmission, and consequently, the mapping function can be a function of the entire state. Similarly, the decoding function can be a function of the fading state. With this queue-architecture, the number of virtual queues at each relay is $|\mathcal{M}||\mathcal{F}|^N$. This is considerably less compared to the number of virtual queues required in the state-based approach, and thus provides a low-complexity queue-architecture. Note that the gain is high in the setting when the number of destinations is large and number of relays is small, which is the

case in cellular systems.

The queue dynamics is as follows: During block t, if the fading state for the first-hop is $\mathbf{g}_1$ and if the central controller decides that the source should transmit a packet using encoding scheme m, then the following queues get updated:

$$Q_S^k[t+1] = (Q_S^k[t] + A^k[t] - r_m^k T)^+, \forall k, \tag{3.8}$$

$$Q_n^{m,\mathbf{g}_1}[t+1] = Q_n^{m,\mathbf{g}_1}[t] + T, \forall n. \tag{3.9}$$

Here, r_m^k denotes the k-th entry in $\mathbf{r}_m$, which is the encoding rate corresponding to destination D_k. During block t, if the fading state for the second-hop is $\mathbf{g}_2$, then the central controller can decide to transmit packets from queues $Q_n^{m,\mathbf{g}_1}, \forall n$ for some given m and $\mathbf{g}_1$ only if $(m, \mathbf{g}_1, \mathbf{g}_2) \in \mathcal{I}$. This ensures that the packet is received successfully at all the destinations. In this case, the following queues get updated:

$$Q_S^k[t+1] = Q_S^k[t] + A^k[t], \forall k, \tag{3.10}$$

$$Q_n^{m,\mathbf{g}_1}[t+1] = (Q_n^{m,\mathbf{g}_1}[t] - T)^+, \forall n. \tag{3.11}$$

Remark 3.3. *In the queue updates (3.8)-(3.9), $r_m^k T$ bits from the source queue Q_S^k is encoded into T symbols and buffered in the relay queue $Q_n^{m,\mathbf{g}_1}$. Note that the source queue lengths $Q_S^k[t]$ are in units of bits whereas relay queue lengths $Q_n^{m,\mathbf{g}_1}[t]$ are in symbols.*

Next, we address the question of designing a central controller that does not require the knowledge of the arrival rates or the fading state distribution.

3.6 Throughput-Optimal Network Algorithm

In this section, we show that a throughput-optimal central[6] controller can be designed without the knowledge of the arrival rates or the fading state distribution. The following algorithm is motivated from back-pressure based algorithms for non-cooperative networks.

Back-pressure-based Algorithm: In every block, the central controller makes decisions based on the current fading state of the system and the current queue-lengths. Let the fading-state during block t be $\mathbf{f}[t] = (\mathbf{f}_1, \mathbf{f}_2)$. The network algorithm run by the controller is as follows:

1. It computes
$$A = \max_{m} \sum_{k} \left(Q_S^k[t] - r_m^k \sum_{n=1}^{N} Q_n^{m,\mathbf{f}_1}[t] \right) r_m^k$$
and an optimal parameter m^* for this problem.

2. It computes
$$\begin{aligned} B = \quad & \max_{m,\mathbf{g}_1} \quad (\mathbf{r}_m \cdot \mathbf{1})^2 \sum_{n=1}^{N} Q_n^{m,\mathbf{g}_1}[t], \\ & \text{s.t.} \quad (m, (\mathbf{g}_1, \mathbf{f}_2)) \in \mathcal{I}, \end{aligned}$$
and a set of optimal parameters $\hat{m}$ and $\hat{\mathbf{g}}_1$ for this problem.

3. If $A \geq B$, then the central controller decides to transmit a packet from the source to the relays using encoding scheme m^*.

[6]Since cooperative schemes require strong node coordination, the centralized nature of the algorithm does not create additional system requirements.

4. Otherwise, the controller decides to transmit a packet from queues $Q_n^{\hat{m},\hat{\mathbf{g}}_1}, \forall n$, i.e., from the relays to the destinations.

The controller repeats steps $1-4$ in every block.

The following theorem provides a strong theoretical guarantee on the throughput performance of this algorithm.

Theorem 3.3. *The above algorithm stochastically stabilizes all the queues for all $\boldsymbol{\lambda}$ such that there exists $\epsilon > 0$ with $\boldsymbol{\lambda} + \epsilon\mathbf{1}$ in the throughput region given in Lemma 3.1, i.e., the underlying network Markov chain is positive recurrent. In simple terms, the algorithm is throughput-optimal.*

Before proceeding to the proof of this theorem, we state the following lemma that is used in the proof.

Lemma 3.4. *Suppose that there exists $\epsilon > 0$ such that $\boldsymbol{\lambda} + \epsilon\mathbf{1}$ is within the throughput region. Then, there exists $a_{\mathbf{f}}^{m,\mathbf{g}} \geq 0$, $b_{\mathbf{f}}^{m,\mathbf{g}} \geq 0$ and $\delta > 0$ such that the following set of conditions are satisfied:*

$$\begin{aligned}
&\lambda_k - \sum_{m,\mathbf{g},\mathbf{f}} (\pi_{\mathbf{f}} r_m^k a_{\mathbf{f}}^{m,\mathbf{g}}) \leq -\delta, \forall k,\\
&\sum_{\mathbf{f}} \pi_{\mathbf{f}} (a_{\mathbf{f}}^{m,\mathbf{g}} - b_{\mathbf{f}}^{m,\mathbf{g}}) \leq -\delta, \forall (m,\mathbf{g},)\\
&\sum_{m,\mathbf{g}} a_{\mathbf{f}}^{m,\mathbf{g}} + b_{\mathbf{f}}^{m,\mathbf{g}} \leq 1, \forall \mathbf{f},\\
&a_{\mathbf{f}}^{m,\mathbf{g}} = 0, \forall (\mathbf{f},\mathbf{g}) \notin \mathcal{I}_1, \forall (m,\mathbf{g}) \notin \mathcal{I},\\
&b_{\mathbf{f}}^{m,\mathbf{g}} = 0, \forall (\mathbf{f},\mathbf{g}) \notin \mathcal{I}_2, \forall (m,\mathbf{g}) \notin \mathcal{I}.
\end{aligned}$$

Proof. Please see Appendix A.2. □

3.6.1 Proof of Theorem 3.3

Since the queues form a Markov chain, we use Foster-Lyapunov theorem in order to prove the stability [9,86]. Without loss of generality, we assume that $\mathbf{r}_m \neq \mathbf{0}, \forall m$. Otherwise, those queues at the relays can be removed without affecting the throughput region and the stability of the system. Now, consider the Lyapunov function

$$V(\mathbf{Q}[t]) = \sum_k \left(Q_s^k[t]\right)^2 + \sum_{n=1}^{N} \sum_{m,\mathbf{g}_1} \left(\mathbf{r}_m \cdot \mathbf{1} Q_n^{m,\mathbf{g}_1}[t]\right)^2,$$

where $\mathbf{Q}[t]$ denotes the vector of all queue lengths.

Next, we consider an optimization problem that captures the algorithm given in this section. Consider a fading-state $\mathbf{f}$ and the following discrete optimization problem:

$$\begin{aligned}
\max_{\alpha_{\mathbf{f}}^{m,\mathbf{g}}, \beta_{\mathbf{f}}^{m,\mathbf{g}}} \quad & \sum_{m,\mathbf{g},k} \left[\left(Q_S^k[t] - r_m^k \sum_{n=1}^{N} Q_n^{m,\mathbf{g}_1}[t] \right) r_m^k \alpha_{\mathbf{f}}^{m,\mathbf{g}} \right] \\
& + \sum_{m,\mathbf{g}} \left[(\mathbf{r}_m \cdot \mathbf{1})^2 \left(\sum_{n=1}^{N} Q_n^{m,\mathbf{g}_1}[t] \right) \beta_{\mathbf{f}}^{m,\mathbf{g}} \right], \qquad (3.12) \\
\text{s.t.} \quad & \sum_{m,\mathbf{g}} (\alpha_{\mathbf{f}}^{m,\mathbf{g}} + \beta_{\mathbf{f}}^{m,\mathbf{g}}) \leq 1, \\
& \alpha_{\mathbf{f}}^{m,\mathbf{g}} = 0, \forall (\mathbf{f}, \mathbf{g}) \notin \mathcal{I}_1, \\
& \beta_{\mathbf{f}}^{m,\mathbf{g}} = 0, \forall (\mathbf{f}, \mathbf{g}) \notin \mathcal{I}_2, \forall (m, \mathbf{g}) \notin \mathcal{I}, \\
& \alpha_{\mathbf{f}}^{m,\mathbf{g}}, \beta_{\mathbf{f}}^{m,\mathbf{g}} \in \{0, 1\}, \forall m, \mathbf{g}.
\end{aligned}$$

Remark 3.4. *It is fairly straightforward to check that the algorithm given in this section results from this optimization problem. Note that this optimization has many redundant variables that are introduced for the purpose of the proof.*

Let an optimal assignment to the optimization problem in (3.12) be $\hat{\alpha}_{\mathbf{f}}^{m,\mathbf{g}}, \hat{\beta}_{\mathbf{f}}^{m,\mathbf{g}}$. Now, from (3.8), (3.10), (3.9) and (3.11), we can bound queue-lenths during block $t+1$ as follows:

$$\begin{aligned}
(Q_S^k[t+1])^2 &= \left(Q_S^k[t] + A^k[t] - \left(\sum_{m,\mathbf{g}} r_m^k T \hat{\alpha}_{\mathbf{f}}^{m,\mathbf{g}}\right)\right)^2 \\
&\leq (Q_s^k[t])^2 + (A^k[t])^2 + \left(\sum_{m,\mathbf{g}} r_m^k T \hat{\alpha}_{\mathbf{f}}^{m,\mathbf{g}}\right)^2 \\
&\quad -2Q_s^k[t]\left(\sum_{m,\mathbf{g}} r_m^k T \hat{\alpha}_{\mathbf{f}}^{m,\mathbf{g}} - A^k[t]\right), \forall k,
\end{aligned}$$

$$\begin{aligned}
(\mathbf{r}_m \cdot \mathbf{1} Q_n^{m,\mathbf{g}_1}[t+1])^2 &\leq \left(\mathbf{r}_m \cdot \mathbf{1} Q_n^{m,\mathbf{g}_1}[t] + \mathbf{r}_m \cdot \mathbf{1} T \sum_{\mathbf{g}_2}\left(\hat{\alpha}_{\mathbf{f}}^{m,\mathbf{g}} - \hat{\beta}_{\mathbf{f}}^{m,\mathbf{g}}\right)\right)^2 \\
&= (\mathbf{r}_m \cdot \mathbf{1} Q_n^{m,\mathbf{g}_1}[t])^2 + \left(\mathbf{r}_m \cdot \mathbf{1} T \sum_{\mathbf{g}_2}\left(\hat{\alpha}_{\mathbf{f}}^{m,\mathbf{g}} - \hat{\beta}_{\mathbf{f}}^{m,\mathbf{g}}\right)\right)^2 \\
&\quad -2(\mathbf{r}_m \cdot \mathbf{1})^2 Q_n^{m,\mathbf{g}_1}[t] T \sum_{\mathbf{g}_2}\left(\hat{\alpha}_{\mathbf{f}}^{m,\mathbf{g}} - \hat{\beta}_{\mathbf{f}}^{m,\mathbf{g}}\right), \forall m, \mathbf{g}_1.
\end{aligned}$$

Applying the law of iterated expectations, we obtain

$$\begin{aligned}
&\mathbf{E}\left[V\left(\mathbf{Q}\left[t+1\right]\right) - V\left(\mathbf{Q}\left[t\right]\right) | \mathbf{Q}\left[t\right]\right] - M \\
&\leq \sum_{\mathbf{f}} \pi_{\mathbf{f}} \left[-\sum_k 2Q_s^k[t]\left(\sum_{m,\mathbf{g}} r_m^k T \hat{\alpha}_{\mathbf{f}}^{m,\mathbf{g}} - \lambda_k T\right) - \right. \\
&\quad \left. \sum_{m,\mathbf{g}_1,n}\left(2(\mathbf{r}_m \cdot \mathbf{1})^2 Q_n^{m,\mathbf{g}_1}[t] T \sum_{\mathbf{g}_2}\left(\hat{\alpha}_{\mathbf{f}}^{m,\mathbf{g}} - \hat{\beta}_{\mathbf{f}}^{m,\mathbf{g}}\right)\right)\right] \\
&= 2T\left[\sum_k Q_S^k[t]\left(\lambda_k - \sum_{m,\mathbf{g},\mathbf{f}}\left(\pi_{\mathbf{f}} r_m^k \hat{\alpha}_{\mathbf{f}}^{m,\mathbf{g}}\right)\right) + \right. \\
&\quad \left. \sum_{m,\mathbf{g},n}(\mathbf{r}_m \cdot \mathbf{1})^2 Q_n^{m,\mathbf{g}_1}[t]\left(\sum_{\mathbf{f}} \pi_{\mathbf{f}}\left(\hat{\alpha}_{\mathbf{f}}^{m,\mathbf{g}} - \hat{\beta}_{\mathbf{f}}^{m,\mathbf{g}}\right)\right)\right]. \qquad (3.13)
\end{aligned}$$

In the last step, we have used the fact that $\sum_{\mathbf{f}} \pi_{\mathbf{f}} = 1$. M is a finite positive value. This follows as variance associated with the arrival processes are bounded and the throughput region is compact.

Let $a_{\mathbf{f}}^{m,\mathbf{g}}, b_{\mathbf{f}}^{m,\mathbf{g}}$ be the values given by Lemma 3.4. Now, substituting values $a_{\mathbf{f}}^{m,\mathbf{g}}$ instead of $\hat{\alpha}_{\mathbf{f}}^{m,\mathbf{g}}$ and $b_{\mathbf{f}}^{m,\mathbf{g}}$ instead of $\hat{\beta}_{\mathbf{f}}^{m,\mathbf{g}}$ in right hand side of (3.13) increases its value. This is due to the following reason. First, consider the linear program (LP) obtained by relaxing the integer constraints of the optimization problem (3.12) and introducing non-negativity constraints. This relaxation is tight as LPs have at least one optimal solution which is a boundary point. Next, the possible values for $a_{\mathbf{f}}^{m,\mathbf{g}}, b_{\mathbf{f}}^{m,\mathbf{g}}$ is a subset of the feasible set for the LP. Therefore, by substituting results from Lemma 3.4 in (3.13), we have

$$\begin{aligned}
&\mathbf{E}\left[V(\mathbf{Q}[t+1]) - V(\mathbf{Q}[t])|\mathbf{Q}[t]\right] - M \\
&\leq \; 2T\left[\sum_k Q_S^k[t]\left(\lambda_k - \sum_{m,\mathbf{g},\mathbf{f}}(\pi_{\mathbf{f}} r_m^k a_{\mathbf{f}}^{m,\mathbf{g}})\right) + \right. \\
&\qquad \left. \sum_{m,\mathbf{g},n}(\mathbf{r}_m \cdot \mathbf{1})^2 Q_n^{m,\mathbf{g}_1}[t]\left(\sum_{\mathbf{f}} \pi_{\mathbf{f}}(a_{\mathbf{f}}^{m,\mathbf{g}} - b_{\mathbf{f}}^{m,\mathbf{g}})\right)\right] \\
&\leq \; -2T\delta\left[\sum_k Q_s^k[t] + \sum_{m,\mathbf{g},n}(\mathbf{r}_m \cdot \mathbf{1})^2 Q_n^{m,\mathbf{g}_1}[t]\right]. \qquad (3.14)
\end{aligned}$$

Now, from (3.14), it is fairly straightforward to see that there is strict negative drift except on a compact subset of the set of queue-lengths. This completes the proof. □

3.6.2 Simulation Results

Consider a two-hop network with a single source, $N = 2$ relays and $K = 10$ destinations. Now, consider correlated ON-OFF fading channels as follows: All the first-hop links are ON or OFF simultaneously with probability p for ON. Similarly, all the second-hop links are ON or OFF simultaneously independent of the first-hop links with probability p for ON. For encoding schemes, we consider the set of cooperative schemes with coding for one destination alone. Hence, the relays would keep virtual queues for K encoding schemes, i.e., one for each destination. Without loss of generality, let $T = 1$, and $r = 1$ bit/symbol be the rate achieved from the source to any destination.

For the above example, the throughput region is as follows:

$$\left\{ \boldsymbol{\lambda} : \lambda_k \geq 0, \forall k, \sum_{k=1}^{K} \lambda_k \leq 0.5p^2 + p(1-p) \right\}.$$

Let the arrival processes corresponding to all the destinations be independent (over processes) i.i.d. Bernoulli with parameter λ. Define $\rho = \lambda K/(0.5p^2 + p(1-p))$ as the load factor. Clearly, the queues will be unstable for load factor $\rho > 1$. We simulate the algorithm presented in this section with channel parameter $p = 0.5$ for load factors $\rho = 0.99$ and $\rho = 1.02$. Simulation trace of one of the queues at the source for 10^5 time slots in plotted in Figure 3.4, and the trace of one of the virtual queues at one of the relays is plotted in Figure 3.5. It is easy to observe that the queues are stable for $\rho = 0.99$ and unstable for $\rho = 1.02$.

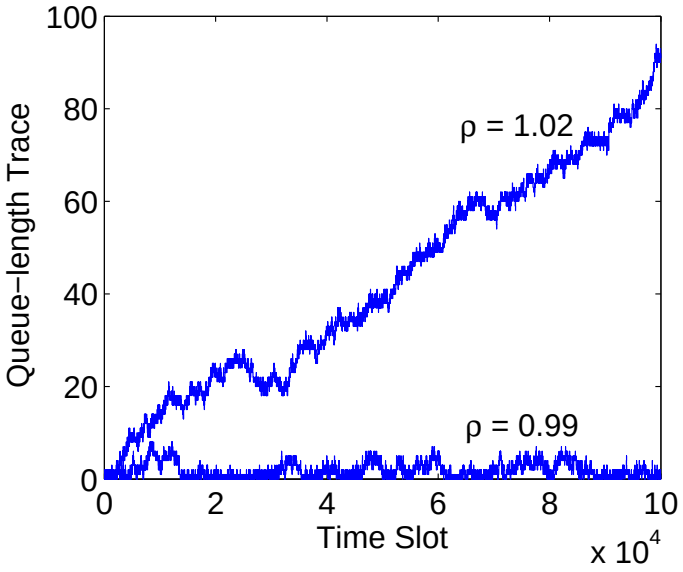

Figure 3.4: Simulation trace of one of the queues at the source

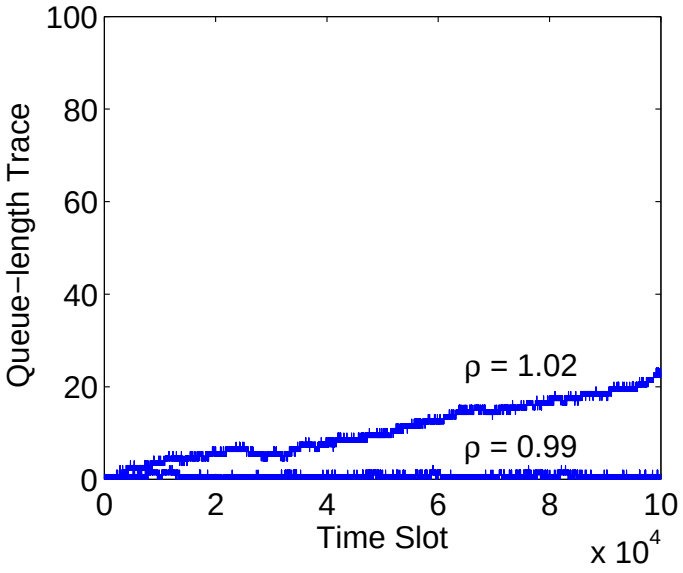

Figure 3.5: Simulation trace of one of the virtual queues at one of the relays

3.6.3 Extensions

So far, we focused on the single-source two-hop scenario because it applies to the heterogeneous cellular network downlink. However, this scenario also captures the essential aspects of scheduling in general cooperative relay network. Hence, the analysis presented will extend to the multi-source multi-hop scenario. The queue-architecture can be applied by extending the state-space, and the back-pressure-based algorithm can be modified in a fairly straightforward manner to account for multi-hop and arbitrary link activation constraints. The complexity of the queue-architecture would increase exponentially with the number of hops. Hence, an area for further study is reducing the complexity of queue-architecture by exploiting specific multi-hop cooperative schemes.

Chapter 4

Distributed Rate Allocation for Wireless Networks

4.1 Introduction

The throughput of wireless networks is traditionally studied separately at the physical and medium access layers, and thus independently optimized at each of these two layers. As a result, conventionally, data-rate adaptation is performed at the physical-layer for each link, and link scheduling is performed at the access-layer. There are significant throughput gains in studying these two in a cross-layer framework [26,39,44,80,120]. This cross-layer optimization results in a joint rate allocation for all the links in the network. Cross-layer approach has been successfully applied to multiple access channels (MAC) to even guarantee delay-optimal rate allocation [134].

Maximum Weighted (Max-Weight) scheduling introduced in the seminal paper [120] performs joint rate allocation and guarantees throughput-optimality.[1] However, Max-Weight algorithm and its variants have the following disadvantages. (*i*) It requires periodic solving of a possibly hard optimization problem. (*ii*) The optimization problem is centralized, and thus

[1] As we already saw in Chapter 3, for cooperative networks, throughput-optimal rate allocation does not follow from classical Max-Weight scheduling.

introduces significant overhead due to queue-length information exchanges. Thus, in order to overcome these disadvantages, we need efficient distributed algorithms for general physical-layer interference models [80].

Our goal is to perform joint rate allocation in a decentralized manner. A related problem is distributed resource allocation in networks, and this problem has received considerable attention in diverse communities over years. In data and/or stochastic processing networks, resource-sharing is typically described in terms of independent set constraints. With such independent set constraints, the resource allocation problem translates to medium access control (or link scheduling) in wireless networks. For such on-off scheduling, recently, efficient algorithms have been proposed for both random access networks [50,117] and CSMA networks [14,83]. More recently, with instantaneous carrier sensing, a throughput-optimal algorithm with local exchange of control messages that approximate Max-Weight has been proposed in [99], and a fully decentralized algorithm has been proposed in [59]. The decentralized queue-length based scheduling algorithm in [59] and its variants have been shown to be throughput-optimal in [57, 58, 81]. This body of literature on completely distributed on-off scheduling has been extended to a framework that incorporates collisions in [60, 91]. Further, this decentralized framework has been validated through experiments in [78].

However, independent set constraints can only model orthogonal channel access which, in general, is known to be sub-optimal [30] (Section 15.1). For wireless networks, the interaction among nodes require a much more fine-

grained characterization than independent set constraints. This can be fully captured in terms of the network's *rate region*, i.e., the set of link-rates that are simultaneously sustainable in the network, for single-hop networks. As long as the data-rates of links are within the *rate region*, simultaneous transmission is possible even by neighboring links in the network. Therefore, it is crucial to perform efficient distributed joint rate allocation (and not just distributed link scheduling) in wireless networks. Although distributed rate allocation is a difficult problem in general, we show that taking advantage of physical-layer information can solve this problem.

4.1.1 Our Contributions

We consider single-hop[2] wireless networks. We present a simple, completely distributed algorithm for rate allocation in wireless networks that is throughput-optimal. In particular, given any rate region for a wireless network, we develop a decentralized (local queue-length based) algorithm that stabilizes all the queues for all arrival rates within the throughput region. Thus, we can utilize the entire physical-layer throughput region of the system with distributed rate allocation. This is an exciting result as the decentralized algorithm achieves the same throughput region as optimal centralized cross-layer algorithms. The algorithm requires that each link can determine the global feasibility of increasing its data-rate from the current data-rate. In Sec-

[2]The results are likely to generalize to multi-hop by combining back-pressure idea with the algorithmic framework developed.

tion 4.8.2, we provide details on techniques to determine rate feasibility, and explain reasons for using this approach in practice.

Our framework builds on the existing distributed link scheduling framework. As discussed before, the current distributed link scheduling algorithms primarily deal with binary (on-off) decisions whereas our algorithm performs scheduling over multiple data-rates. Similar to these existing distributed link scheduling algorithms, our algorithm is mathematically modeled by a Markov process on the discrete set of data-rates. However, with multiple data-rates for each link, the appropriate choice of the large number of transition rates is complicated. Thus, a key challenge is to design a Markov chain with fewer parameters that can be analyzed and appropriately chosen for throughput-optimality. We overcome this challenge by showing that transition rates with the following structure have this property. For link i, the transition rate to a data-rate $r_{i,j}$ from any other data-rate can be set to $\exp(r_{i,j}v_i)$, where v_i is a single parameter associated with link i that is updated based on its queue-length. The transition takes place only if the new data-rate is feasible. As expected, this reduces to the existing algorithmic framework in the special case of binary (on-off) decisions. Thus, our main contributions are:

- We generalize existing on-off framework to multi-state framework.
- We utilize this framework to develop distributed rate allocation algorithms for wireless networks that are (i) throughput-optimal, and (ii) completely decentralized.

For the general framework mentioned above, at an intuitive level, the techniques required for proving throughput-optimality remain similar to existing techniques. However, there are few additional technical issues that arise while analyzing the general framework. First, we need to account for more general constraints that arise from the set of possible rate allocation vectors. Next, the choice of update rules for $v_i(t)$ with time t based on local queue-lengths that guarantee throughput-optimality does not follow directly. The mixing time of the rate allocation Markov chain plays an important role in choosing the update rules. For arbitrary throughput regions, any rate allocation algorithm that approach ϵ-close (for arbitrarily small ϵ) to the boundary possibly requires an increasing $1/\epsilon$ number of data-rates per link. This leads to a potential increase in the mixing time due to the increase in the size of the state-space. Thus, the analysis performed is more general and essential to establish throughput-optimality of the algorithms considered.

An important application of this algorithmic framework is for networks of white-space radios [34], where multiple non-adjacent frequency bands are available for operation and multiple radios are available at the wireless nodes. A scheduler needs to allocate different radios to different bands in a distributed manner. This problem introduces multiple data-rates for every link even in the traditional *carrier sensing*[3] framework, and hence, existing distributed algorithms cannot be directly applied. We demonstrate that our framework

[3] Note that this application uses traditional carrier sensing and does not require any additional feasibility assumption. Thus, our framework is of importance in practice irrespective of the feasibility assumption.

provides a throughput-optimal distributed algorithm in this setting.

4.1.2 Organization

In Section 4.2, we present the system model. In Section 4.3, we introduce relevant definitions and known results used. In Section 4.4, we explain the distributed rate allocation algorithm and state the main results. In Section 4.5, we describe the rate allocation Markov chain and the optimization framework. In Section 4.6, we establish the throughput-optimality of the adaptive algorithm. The algorithm for multiple-band multiple-radio scheduling is given in Section 4.7. Further discussions and simulation results are given in Section 4.8. The proofs of the technical lemmas in Section 4.5 and Section 4.6 are given in Appendix B.

4.2 System Model

Consider a wireless network consisting of m nodes, labeled $\mathcal{N}$. In this network, we are interested in n single-hop flows that correspond to n wireless links labeled $\mathcal{L}$. Since we have a shared wireless medium, these links interact (or interfere) in a potentially complex way. For single-hop flows, this interaction among links can be captured through a n-dimensional *rate region* for the network, which is formally defined next.

Definition 4.1 (Rate Region)**.** *The rate region of a network, denoted by* $\mathcal{C} \subseteq \mathbb{R}_+^n$*, is defined as the set of instantaneous rate vectors* $\mathbf{c} \in \mathbb{R}_+^n$ *at which queues (introduced later) of all* n *links can be drained simultaneously.*

We assume that the rate region is fixed[4] (i.e., not time-varying). By definition, this rate region is compact. We assume that the rate region has the following simple property: if $\mathbf{c} \in \mathcal{C}$, then $\hat{\mathbf{c}} \in \mathcal{C}$ for all $\hat{\mathbf{c}} \leq \mathbf{c}$ and $\hat{\mathbf{c}} \geq \mathbf{0}$. The above property states that rates can be decreased component-wise. Such an assumption is fairly mild, and is satisfied by rate regions resulting from most physical-layer schemes.

Next, we define the throughput region of the network.

Definition 4.2 (Throughput Region)**.** *The throughput region of a network, denoted by $\mathcal{T}$, is defined as the convex hull of the rate region $\mathcal{C}$ of the network.*

We use a continuous-time model to describe system dynamics. Time is denoted by $t \in \mathbb{R}_+$. Every (transmitter of) link $i \in \mathcal{L}$ is associated with a queue $Q_i(t) \in \mathbb{R}_+$, which quantifies the information (packets) remaining at time t waiting to be transmitted on link i. Let the cumulative arrival of information at the i-th link during the time interval $[0, t)$ be $A_i(t) \in \mathbb{R}_+$ with $A_i(0) := 0$. *Rate allocation* at time t is defined as the rate vector in the rate region at which the system is being operated at time t. Let the rate allocation corresponding to the i-th link at time t be $r_i(t)$. Then, for every link $i \in \mathcal{L}$, the queue dynamics is given by

$$Q_i(t) = Q_i(s) - \int_s^t r_i(z)\mathbf{I}(Q_i(z) > 0)dz + A_i(t) - A_i(s), \tag{4.1}$$

where $0 \leq s < t$. The vector of n queues in the system is denoted by $\mathbf{Q}(t) = [Q_i(t)]_{i=1}^n$. The queues are initially at $\mathbf{Q}(0) \in \mathbb{R}_+^n$.

[4]We consider fixed or slow-fading channels.

We consider arrival processes at the queues in the network with the following properties.

- We assume every arrival process is such that increments over integral times are independent and identically distributed with $\Pr(A_i(1) = 0) > 0$.
- We assume that all these increments belong to a bounded support $[0, K]$, i.e., $A_i(k+1) - A_i(k) \in [0, K]$ for all $k \in \mathbb{Z}_+$.

Based on these properties, the (mean) *arrival rate* corresponding to the i-th link is $\lambda_i := \mathbb{E}[A_i(1)]$. We denote the vector of arrival rates by $\boldsymbol{\lambda}$. Without loss of generality,[5] we assume $\lambda_{\min} := \min_i \lambda_i > 0$. It follows from the strong law of large numbers that, with probability 1,

$$\lim_{t\to\infty} \frac{A_i(t)}{t} = \lambda_i. \tag{4.2}$$

In summary, our system model incorporates general interference constraints through an arbitrary rate region and focuses on single-hop flows.

4.3 Definitions & Known Results

We provide definitions and known results that are key in establishing the main results. We begin with definitions on two measures of difference between two probability distributions.

[5] If $\lambda_i = 0$, then this link can be removed from the system.

Definition 4.3 (Kullback-Leibler (KL) divergence)**.** *Consider two probability mass functions π and μ on a finite set $\mathcal{X}$. Then, the KL divergence from π to μ is defined as $D(\mu\|\pi) = \sum_{\mathbf{x}\in\mathcal{X}} \mu(\mathbf{x}) \log \frac{\mu(\mathbf{x})}{\pi(\mathbf{x})}$.*

Definition 4.4 (Total Variation)**.** *Consider two probability mass functions π and μ on a finite set $\mathcal{X}$. Then, the total variation distance between π and μ is defined as $\|\mu - \pi\|_{TV} = \frac{1}{2}\sum_{\mathbf{x}\in\mathcal{X}} |\mu(\mathbf{x}) - \pi(\mathbf{x})|$.*

Next, we provide two known results that are used later. Result 4.1 follows directly from [17](Theorem 3.2), and Result 4.2 is in [17](Theorem 4.3).

Result 4.1 (Mixing Time)**.** *Consider any finite state-space, aperiodic, irreducible, discrete-time Markov chain with transition probability matrix P and the stationary distribution $\boldsymbol{\alpha}$. Let $\alpha_{\min}$ be the minimum value in $\boldsymbol{\alpha}$ and the second largest eigenvalue modulus (SLEM) be $\sigma_{\max}$. Then, for any $\rho > 0$, starting from any initial distribution (at time 0), the distribution at time $\tau \in \mathbb{Z}_{++}$ associated with the Markov chain, denoted by $\boldsymbol{\beta}(\tau)$, is such that $\|\boldsymbol{\beta}(\tau) - \boldsymbol{\alpha}\|_{TV} \leq \rho$ if*

$$\tau \geq \frac{\frac{1}{2}\log(1/\alpha_{\min}) + \log(1/\rho)}{\log(1/\sigma_{\max})}. \tag{4.3}$$

Result 4.2 (Conductance Bounds)**.** *Consider the setting as above. The ergodic flow out of $\mathcal{S} \subseteq \mathcal{X}$ is defined as $F(\mathcal{S}) = \sum_{\mathbf{x}\in\mathcal{S},\hat{\mathbf{x}}\in\mathcal{S}^c} \boldsymbol{\alpha}(\mathbf{x})P(\mathbf{x},\hat{\mathbf{x}})$ and the conductance is defined as*

$$\Phi = \min\left\{ \frac{F(\mathcal{S})}{\sum_{\mathbf{x}\in\mathcal{S}} \boldsymbol{\alpha}(\mathbf{x})} : \phi \subset \mathcal{S} \subset \mathcal{X}, \sum_{\mathbf{x}\in\mathcal{S}} \boldsymbol{\alpha}(\mathbf{x}) \leq \frac{1}{2} \right\}. \tag{4.4}$$

Then, the SLEM $\sigma_{\max}$ is bounded by conductance as follows:

$$1 - 2\Phi \leq \sigma_{\max} \leq 1 - \Phi^2/2. \tag{4.5}$$

Lastly, we provide the definition of positive Harris recurrence. For details on properties associated with positive Harris recurrence, see [33, 86].

Definition 4.5 (Positive Harris recurrence)**.** *Consider a discrete-time time-homogeneous Markov chain on a complete, separable metric space X. Let $\mathcal{B}_{\mathcal{X}}$ denote the Borel σ-algebra on X, and X_τ denote the state of the Markov chain at time $\tau \in \mathbb{Z}_+$. Define stopping time $T_A = \inf\{\tau \geq 0 : X_\tau \in A\}$ for any $A \in \mathcal{B}_{\mathcal{X}}$. The set A is called Harris recurrent if $\Pr(T_A < \infty | X(0) = x) = 1$ for any $x \in X$. A Markov chain is called Harris recurrent if there exits a σ-finite measure μ on $(X, \mathcal{B}_{\mathcal{X}})$ such that if $\mu(A) > 0$ for some $A \in \mathcal{B}_{\mathcal{X}}$, then A is Harris recurrent. It is known that if X is Harris recurrent an essentially unique invariant measure exists. If the invariant measure is finite, then it may be normalized to a probability measure. In this case, X is called positive Harris recurrent.*

We proceed to describe the rate allocation algorithm and our main results.

4.4 Rate Allocation Algorithm & Main Results

Our goal is to design a completely decentralized algorithm for rate allocation that *stabilizes* all the queues as long as the arrival rate vector is

within the throughput region. By assumption, every link can determine rate feasibility, i.e., every link can determine whether increasing its data-rate from the current rate allocation results in a net feasible rate vector. More formally, every link $i \in \mathcal{L}$ at time t, if required, can obtain the information $\mathbf{I}(\mathbf{r}(t)+\alpha\mathbf{e}_i \in \mathcal{C})$, for any $\alpha \in \mathbb{R}$. More details on determining rate feasibility are given in Section 4.8.

The rate allocation vector at time t is denoted by $\mathbf{r}(t) = [r_i(t)]_{i=1}^n$. For **decentralized rate allocation**, we develop an algorithm that uses only local queue information for choosing $\mathbf{r}(t)$ over time t. Further, we perform rate allocation over a chosen limited (finite) set of rate vectors that are *feasible*. We choose a finite set of rate levels corresponding to every link, and form vectors that are feasible. The details are as follows:

1. For each link $i \in \mathcal{L}$, a set of rate levels $\mathcal{R}_i = \{r_{i,j}\}_{j=0}^{k_i}$ are chosen from $[0, c_i]$ with $r_{i,0} = 0$, $r_{i,k_i} = c_i$ and $r_{i,j-1} < r_{i,j}$. Here, c_i is the maximum possible transmission rate for the i-th link, i.e., $c_i = \arg\max_{\alpha\in\mathbb{R}_+} \alpha\mathbf{e}_i \in \mathcal{C}$, and $k_i \in \mathbb{Z}_{++}$ is the number of levels other than zero. Since the rate region is compact, without loss of generality,[6] we assume $0 < \underline{K} \leq c_i \leq \bar{K} < \infty$.

2. The set of rate allocation vectors, denoted by $\mathcal{R}$, is given by $\mathcal{R} = \{[r_1, r_2, \ldots, r_n] : r_i \in \mathcal{R}_i \text{ for all } i \in \mathcal{L}, \text{ and } [r_1, r_2, \ldots, r_n] \in \mathcal{C}\}$.

[6] If $c_i = 0$, then this link can be removed from the system.

The convex hull of the set of rate allocation vectors $\mathcal{R}$ is denoted by $\mathcal{R}_c$. Define $\mathcal{R}_c^o = \{\mathbf{r} \in \mathbb{R}_+^n : \mathbf{r} < \mathbf{t} \text{ for some } \mathbf{t} \in \mathcal{R}_c\}$, the set of *strictly* feasible rates. For rate regions that are polytopes, the partitions $\mathcal{R}_i$ can be chosen such that $\mathcal{R}_c = \mathcal{T}$. For any compact rate region, it is fairly straightforward to choose partitions $\mathcal{R}_i$ with $k_i \leq \lceil 2c_i/\epsilon \rceil \leq \left\lceil 2\bar{K}/\epsilon \right\rceil$ such that $\mathbf{c} \in \mathcal{R}_c$ if $\mathbf{c} + \frac{\epsilon}{2}\mathbf{1} \in \mathcal{T}$. The trivial partition with $\epsilon/2$ as step size in all dimensions satisfy the above property. Thus, for any given $\epsilon > 0$, we can obtain a set of rate allocation vectors $\mathcal{R}$ such that

$$|\mathcal{R}| \leq \left\lceil 2\bar{K}/\epsilon \right\rceil^n \tag{4.6}$$

and $\mathbf{c} \in \mathcal{R}_c$ if $\mathbf{c} + \frac{\epsilon}{2}\mathbf{1} \in \mathcal{T}$.

Before describing the algorithm, we define two notions of throughput performance of a rate allocation algorithm.

Definition 4.6 (Rate stable)**.** *We say that a rate allocation algorithm is rate-stable if, for any $\boldsymbol{\lambda} \in \mathcal{R}_c^o$, the departure rate corresponding to every queue is equal to its arrival rate, i.e., for all $i \in \mathcal{L}$, with probability* 1*,*

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t r_i(z)\mathbf{I}(Q_i(z) > 0)dz = \lambda_i.$$

From (4.1),(4.2), this is same as, for all $i \in \mathcal{L}$, with probability 1*,*

$$\lim_{t\to\infty} Q_i(t)/t = 0.$$

Definition 4.7 (Throughput optimal)**.** *We say that a rate allocation algorithm is throughput-optimal if, for any given $\epsilon > 0$, the algorithm makes the*

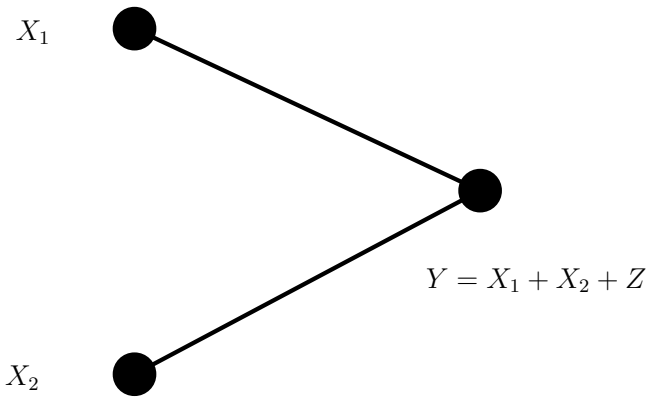

Figure 4.1: Gaussian multiple access channel

underlying network Markov chain positive Harris recurrent *(defined in Section 4.3) for all* $\boldsymbol{\lambda}$ *such that* $\boldsymbol{\lambda} + \epsilon \mathbf{1} \in \mathcal{T}$. *By definition, the algorithm can depend on the value of* ϵ.

Next, we describe **a class of algorithms** to determine $\mathbf{r}(t)$ as a function of time based on a continuous-time Markov chain. Recall that $\mathcal{R}_i = \{r_{i,j}\}_{j=0}^{k_i}$ is the set of possible rates/states for allocation associated with the i-th link. In these algorithms, the i-th link uses k_i independent exponential clocks with rates/parameters[7] $\{U_{i,j}\}_{j=0}^{k_i}$ (or equivalently exponential clocks with mean times $\{1/U_{i,j}\}_{j=0}^{k_i}$). The clock with (time varying) parameter $U_{i,j}$ is associated with the state $r_{i,j}$. Based on these clocks, the i-th link obtains $r_i(t)$ as follows:

1. If the clock associated with a state (say $j = m$) ticks and further if transitioning to that state $r_{i,m}$ is feasible, then $r_i(t)$ is changed to $r_{i,m}$;

[7]These should not to be confused with the rates for allocation.

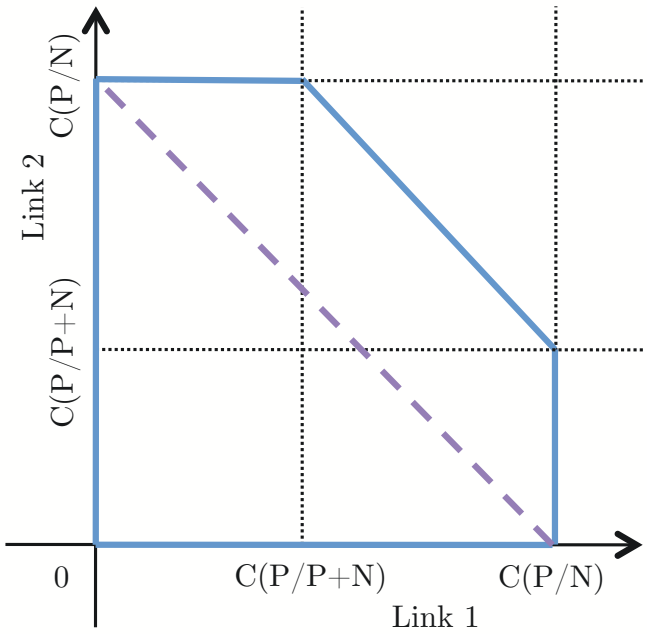

Figure 4.2: Information-theoretic capacity region

2. Otherwise, $r_i(t)$ remains the same.

The above procedure continues, i.e, all the clocks run continuously. Define $u_{i,j} = \log U_{i,j}, \forall i \in \mathcal{L}, j \in \{0, 1, \ldots, k_i\}$. It turns out that the appropriate structure to introduce is as follows:

$$u_{i,j} = r_{i,j} v_i, \forall i \in \mathcal{L}, j \in \{0, 1, \ldots, k_i\},$$

where $v_i \in \mathbb{R}, \forall i \in \mathcal{L}$. We denote the vector consisting of these new set of parameters by $\mathbf{v} = [v_i]_{i=1}^n$.

Example 4.1. *Consider a Gaussian multiple access channel with two links as shown in Figure 4.1 with average power constraint P at the transmitters*

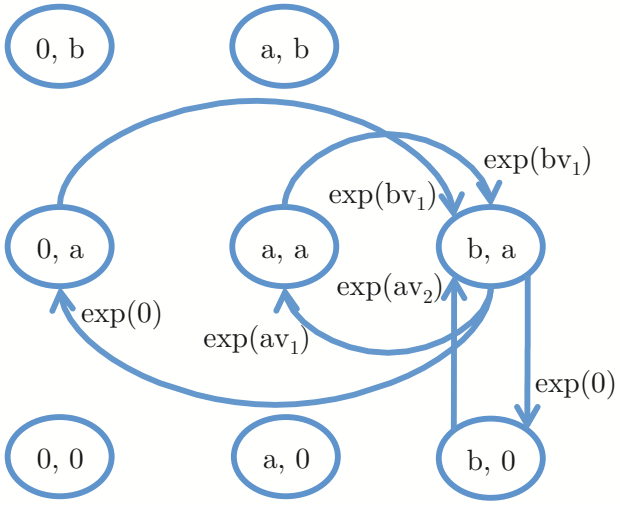

Figure 4.3: Rate allocation Markov chain (transitions to/from (b, a) state alone shown)

and noise variance N at the receiver. The capacity region of this channel is shown in Figure 4.2 where $C(x) = 0.5\log_2(1+x)$. In this case, orthogonal access schemes limit the throughput region to the triangle (strictly within the pentagon) shown using dash-line. In this example, if we allow for capacity-achieving physical-layer schemes, the rate region (and hence the throughput region) is identical to the pentagon shown in Figure 4.2. The natural choice for the set of rate levels at link-1 is $\mathcal{R}_1 = \{0, a, b\}$ where $a = C(P/P+N)$ and $b = C(P/N)$. Similarly, $\mathcal{R}_2 = \{0, a, b\}$. This leads to the set of rate allocation vectors $\mathcal{R} = \{[0,0], [0,a], [0,b], [a,0], [a,a], [a,b], [b,0], [b,a]\}$. It is clear that the convex combination of this set is the throughput region itself. For this example, the state-space of the Markov chain and transitions to and from state (b, a) are shown in Figure 4.3.

A distributed algorithm needs to choose the parameters $\mathbf{v}$ in a decentralized manner. For providing the intuition behind the algorithm, we perform this in two steps. In the first step, we develop the non-adaptive version of the algorithm that has the knowledge of $\boldsymbol{\lambda}$. This algorithm is called non-adaptive as it requires the explicit knowledge of $\boldsymbol{\lambda}$. Despite the fact that it assumes $\boldsymbol{\lambda}$, this algorithm is non-trivial as it does not need the knowledge of instantaneous network-wide queue-lengths. In the second step, we develop the adaptive algorithm, where $\mathbf{v}$ is obtained as a function of time t denoted by $\mathbf{v}(t)$.[8] This algorithm is called adaptive as it does not even require the knowledge of $\boldsymbol{\lambda}$.

4.4.1 Non-Adaptive Algorithm

The rate allocation at time $t = 0$ is set to be $\mathbf{r}(0) = \mathbf{0}$. This algorithm uses $\mathbf{v}^*$ at all times which is a function of $\boldsymbol{\lambda}$, and is given by

$$\mathbf{v}^* = \arg\max_{\mathbf{v}\in\mathbb{R}^n} \boldsymbol{\lambda}\cdot\mathbf{v} - \log\left(\sum_{\mathbf{r}\in\mathcal{R}} \exp\left(\mathbf{r}\cdot\mathbf{v}\right)\right).$$

We show in Section 4.5 that, given $\boldsymbol{\lambda} \in \mathcal{R}_c^o$, the above optimization problem has a unique solution that is finite, and therefore has a valid $\mathbf{v}^*$. An important result regarding this non-adaptive algorithm is the following theorem.

Theorem 4.1. *The above non-adaptive algorithm is rate-stable for any given* $\boldsymbol{\lambda} \in \mathcal{R}_c^o$.

[8] This implies that the exponential clocks used have time varying rates. These are well-defined non-homogeneous Poisson processes.

Proof Outline. For any $\boldsymbol{\lambda} \in \mathcal{R}_c^o$, there is at least one distribution on $\mathcal{R}$ that has expectation as $\boldsymbol{\lambda}$. For the Markov chain specified by any $\mathbf{v} \in \mathbb{R}^n$, there is a stationary distribution on the state-space $\mathcal{R}$. The value $\mathbf{v}^*$ is chosen such that it minimizes the Kullback-Leibler divergence of the induced stationary distribution from the distribution corresponding to $\boldsymbol{\lambda}$. For the Markov chain specified by $\mathbf{v}^*$, the expected value of the stationary distribution turns out to be $\boldsymbol{\lambda}$. This leads to rate-stable performance of the algorithm. The proof details are given in Section 4.5. □

4.4.2 Adaptive Algorithm

The values of $\mathbf{v}(t)$ are updated during fixed (not random variables) time instances τ_l for $l \in \mathbb{Z}_{++}$. We set $\tau_0 = 0$ and $\mathbf{v}(0) = \mathbf{0}$. During interval $t \in [\tau_l, \tau_{l+1})$ the algorithm uses $\mathbf{v}(t) = \mathbf{v}(\tau_l)$. The length of the intervals are $T_l = \tau_{l+1} - \tau_l$. During interval $[\tau_l, \tau_{l+1})$, let the *empirical arrival rate* be

$$\hat{\lambda}_i(l) = \frac{A_i(\tau_{l+1}) - A_i(\tau_l)}{T_l} \tag{4.7}$$

and the *empirical offered service rate* be

$$\hat{s}_i(l) = \frac{1}{T_l} \int_{\tau_l}^{\tau_{l+1}} r_i(z) dz. \tag{4.8}$$

The update equation corresponding to the algorithm for the i-th link is given by

$$v_i(\tau_{l+1}) = \left[v_i(\tau_l) + \alpha_l \left(\hat{\lambda}_i(l) + \frac{\epsilon}{4} - \hat{s}_i(l) \right) \right]_D \tag{4.9}$$

where $[\theta]_D = \min(\theta, D)\mathbf{I}(\theta \geq 0) + \max(\theta, -D)\mathbf{I}(\theta < 0)$, i.e., $[\theta]_D$ is the projection of θ to the closest point in $[-D, D]$, and α_l are the step sizes. Thus, the algorithm parameters are interval lengths T_l, step sizes α_l and D.

Remark 4.1. *Clearly, both empirical arrival rate and empirical offered service rate used in the above algorithm can be computed by the i-th link without any external information. In fact, the difference is simply the difference of its queue-length over the previous interval appropriately scaled by the inverse of the length of the previous interval.*

The following theorem provides $(1-\epsilon)$-optimal performance guarantee for the adaptive algorithm.

Theorem 4.2. *Consider any given $\epsilon > 0$, $\epsilon \leq 4\lambda_{\min}$. Then, there exists some choice of algorithm parameters $T_l = T(n, \epsilon)$, $\alpha_l = \alpha(n, \epsilon)$ and $D = D(n, \epsilon)$ such that the appropriate network Markov chain under the adaptive algorithm is positive Harris recurrent if $\boldsymbol{\lambda} + \epsilon\mathbf{1} \in \mathcal{T}$, i.e., the algorithm is throughput-optimal.*

Proof Outline. The update in (4.9) can be intuitively thought of as a gradient decent technique to solve an optimization problem that will lead to $\mathbf{v}^*$ whose induced stationary distribution on $\mathcal{R}$ has expected value *strictly* greater than $\boldsymbol{\lambda}$. However, the arrival rate and offered service rate are replaced with their empirical values for decentralized operation. We consider the two time scales involved in the algorithm - update interval T and N update intervals.

The main steps involved in establishing the throughput-optimality are the following. First, we show that, sufficiently long T can be chosen such that the empirical values used in the algorithm are arbitrarily close to the true values. Using this, we next show that the average offered empirical service rate over N update intervals is *strictly* higher than the arrival rate. Finally, we show that this results in a *drift* that is sufficient to guarantee positive Harris recurrence. The proof details are given in Section 4.6. □

4.5 Rate Allocation Markov Chain & Rate Stability of the Non-Adaptive Algorithm

Rate allocation Markov chain: The main challenge is to design a Markov chain with fewer parameters that can be analyzed and appropriately chosen for throughput-optimality. First, we identify a class of Markov chains that are relatively easy to analyze. Consider the class of algorithms introduced in Section 4.4. The underlying mathematical object for this class of algorithms is a continuous-time Markov chain with state-space $\mathcal{R}$, which is the (finite) set of rate allocation vectors. Define

$$f(\hat{\mathbf{r}}, \mathbf{r}) = \exp\left(\sum_{i=1}^{n}\sum_{j=0}^{k_i} u_{i,j}\mathbf{I}(r_i = r_{i,j})\mathbf{I}(r_i \neq \hat{r}_i)\right), \tag{4.10}$$

where $\hat{\mathbf{r}} = [\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_n] \in \mathcal{R}$, $\mathbf{r} = [r_1, r_2, \ldots, r_n] \in \mathcal{R}$ and $u_{i,j}$ are the parameters introduced in Section 4.4. Now, the transition rate from state $\hat{\mathbf{r}} \in \mathcal{R}$ to state $\mathbf{r} \in \mathcal{R}$ can be expressed as

$$q(\hat{\mathbf{r}}, \mathbf{r}) = \begin{cases} f(\hat{\mathbf{r}}, \mathbf{r}), & \text{if } \|\hat{\mathbf{r}} - \mathbf{r}\|_0 = 1, \\ 0, & \text{if } \|\hat{\mathbf{r}} - \mathbf{r}\|_0 > 1. \end{cases}$$

And, the diagonal elements of the rate matrix are given by

$$q(\hat{\mathbf{r}}, \hat{\mathbf{r}}) = -\sum_{\mathbf{r}\in\mathcal{R}, \mathbf{r}\neq\hat{\mathbf{r}}} q(\hat{\mathbf{r}}, \mathbf{r}), \forall \hat{\mathbf{r}} \in \mathcal{R}.$$

This follows directly from the description of the algorithm. This class of algorithms are carefully designed such that it is tractable for analysis. In particular, the following lemma shows that this Markov chain is reversible and the stationary distribution has exponential form.

Lemma 4.3. *The rate allocation Markov chain* $(\mathcal{R}, q)$ *is reversible and has the stationary distribution*

$$\pi(\mathbf{r}) = \frac{\exp\left(\sum_{i=1}^{n}\sum_{j=0}^{k_i} u_{i,j}\mathbf{I}(r_i = r_{i,j})\right)}{\sum_{\tilde{\mathbf{r}}\in\mathcal{R}}\exp\left(\sum_{i=1}^{n}\sum_{j=0}^{k_i} u_{i,j}\mathbf{I}(\tilde{r}_i = r_{i,j})\right)}. \tag{4.11}$$

Furthermore, this Markov chain converges to this stationary distribution starting from any initial distribution.

Proof. The proof follows from detailed balance equations

$$\pi(\mathbf{r})q(\mathbf{r}, \hat{\mathbf{r}}) = \pi(\hat{\mathbf{r}})q(\hat{\mathbf{r}}, \mathbf{r})$$

for all $\mathbf{r}, \hat{\mathbf{r}} \in \mathcal{R}$ and known results on convergence to stationary distribution for irreducible finite state-space continuous-time Markov chains [9]. □

The *offered service rate* vector under the stationary distribution is

$$\mathbf{s} := \sum_{\mathbf{r}\in\mathcal{R}} \pi(\mathbf{r})\mathbf{r}.$$

In general, for $\boldsymbol{\lambda} \in \mathcal{R}_c^o$, we expect to find values for parameters $u_{i,j}$ as a function of $\boldsymbol{\lambda}$ and $\mathcal{R}$ such that $\mathbf{s} = \boldsymbol{\lambda}$. Due to the exponential form in (4.11), it turns out that the right structure to introduce is

$$u_{i,j} = r_{i,j} v_i, \forall i \in \mathcal{L}, j \in \{0, 1, \ldots, k_i\}, \tag{4.12}$$

where $v_i \in \mathbb{R}, \forall i \in \mathcal{L}$, and obtain suitable values for $\mathbf{v} = [v_i]_{i=1}^n$ as a function of $\boldsymbol{\lambda}$ and $\mathcal{R}$ such that $\mathbf{s} = \boldsymbol{\lambda}$. To emphasize the dependency on $\mathbf{v}$, from now onwards, we denote the stationary distribution by $\pi_{\mathbf{v}}(\mathbf{r})$ and the offered service rate vector by

$$\mathbf{s}_{\mathbf{v}} = \sum_{\mathbf{r} \in \mathcal{R}} \pi_{\mathbf{v}}(\mathbf{r}) \mathbf{r}. \tag{4.13}$$

Substituting (4.12), we can simplify (4.11) to obtain

$$\pi_{\mathbf{v}}(\mathbf{r}) = \frac{\exp(\mathbf{r} \cdot \mathbf{v})}{\sum_{\tilde{\mathbf{r}} \in \mathcal{R}} \exp(\tilde{\mathbf{r}} \cdot \mathbf{v})}. \tag{4.14}$$

Optimization framework: We utilize the optimization framework in [59] to show that values for $\mathbf{v}$ exist such that $\mathbf{s}_{\mathbf{v}} = \boldsymbol{\lambda}$. In particular, we show that the unique solution to an optimization problem given by $\mathbf{v}^*$ has the property $\mathbf{s}_{\mathbf{v}^*} = \boldsymbol{\lambda}$. Next, we describe the intuitive steps to arrive at the optimization problem. If $\boldsymbol{\lambda} \in \mathcal{R}_c^o$, then $\boldsymbol{\lambda}$ can be expressed as a convex combination of $\mathbf{r} \in \mathcal{R}$, i.e., there exists a valid probability distribution $\mu(\mathbf{r})$ such that $\boldsymbol{\lambda} = \sum_{\mathbf{r} \in \mathcal{R}} \mu(\mathbf{r}) \mathbf{r}$. For a given distribution $\mu(\mathbf{r})$, we are interested in choosing $\mathbf{v}$ such that $\pi_{\mathbf{v}}(\mathbf{r})$ is *close* to $\mu(\mathbf{r})$. We consider the KL divergence of $\pi_{\mathbf{v}}(\mathbf{r})$ from $\mu(\mathbf{r})$ given by $D\left(\mu(\mathbf{r}) \| \pi_{\mathbf{v}}(\mathbf{r})\right)$. Minimizing $D\left(\mu(\mathbf{r}) \| \pi_{\mathbf{v}}(\mathbf{r})\right)$ over the

parameter $\mathbf{v}$ is equivalent in terms of the optimal solution(s) to maximizing $F(\mu(\mathbf{r}), \pi_{\mathbf{v}}(\mathbf{r})) = -D\left(\mu(\mathbf{r}) \| \pi_{\mathbf{v}}(\mathbf{r})\right) - H(\mu(\mathbf{r}))$ over the parameter $\mathbf{v}$ as $H(\mu(\mathbf{r}))$ is a constant. Simplifying $F(\mu(\mathbf{r}), \pi_{\mathbf{v}}(\mathbf{r}))$ leads the optimization problem as follows:

$$\begin{aligned} F(\mu(\mathbf{r}), \pi_{\mathbf{v}}(\mathbf{r})) &= \sum_{\mathbf{r}\in\mathcal{R}} \mu(\mathbf{r}) \log \pi_{\mathbf{v}}(\mathbf{r}) \\ &\overset{(a)}{=} \sum_{\mathbf{r}\in\mathcal{R}} \mu(\mathbf{r})\mathbf{r}\cdot\mathbf{v} - \log\left(\sum_{\mathbf{r}\in\mathcal{R}} \exp(\mathbf{r}\cdot\mathbf{v})\right) \\ &\overset{(b)}{=} \boldsymbol{\lambda}\cdot\mathbf{v} - \log\left(\sum_{\mathbf{r}\in\mathcal{R}} \exp(\mathbf{r}\cdot\mathbf{v})\right). \end{aligned}$$

Here, (a) follows from (4.14) and (b) follows from the assumption

$$\boldsymbol{\lambda} = \sum_{\mathbf{r}\in\mathcal{R}} \mu(\mathbf{r})\mathbf{r}.$$

Now onwards, we denote the objective function by $F(\mathbf{v}, \boldsymbol{\lambda})$. To summarize, the optimization problem of interest is, given $\boldsymbol{\lambda} \in \mathcal{R}_c^o$,

$$\begin{aligned} \text{maximize} \quad & F(\mathbf{v}, \boldsymbol{\lambda}) = \boldsymbol{\lambda}\cdot\mathbf{v} - \log\left(\sum_{\mathbf{r}\in\mathcal{R}} \exp(\mathbf{r}\cdot\mathbf{v})\right) \\ \text{subject to} \quad & \mathbf{v} \in \mathbb{R}^n. \end{aligned} \tag{4.15}$$

The following lemma regarding the optimization problem in (4.15) is a key ingredient to the main results.

Lemma 4.4. *Let $\boldsymbol{\lambda} \in \mathcal{R}_c^o$. The optimization problem in (4.15) has a unique solution $\mathbf{v}^*(\boldsymbol{\lambda})$, which is finite. In addition, the offered service rate vector under $\mathbf{v}^*$ is equal to the arrival rate vector, i.e., $\mathbf{s}_{\mathbf{v}^*} = \boldsymbol{\lambda}$.*

Proof. Please see Appendix B.1. □

The important observations are that the objective function is concave in $\mathbf{v}$ and the gradient with respect to $\mathbf{v}$ is $\boldsymbol{\lambda} - \mathbf{s}_{\mathbf{v}}$. With offered service rate equal to arrival rate, the next step is to show that the queues drain at rate equal to $\boldsymbol{\lambda}$.

4.5.1 Proof of Theorem 4.1

Rate stability of the non-adaptive algorithm: Consider time instances ν_l for $l \in \mathbb{Z}_+$ with $\nu_0 = 0$, and interval length $\Gamma_l = \nu_{l+1} - \nu_l = l + 1$. The queue at the i-th link can be upper bounded as follows. The offered service during the time interval is $[\nu_k, \nu_{k+1})$ is used to serve the arrivals during the time interval $[\nu_{k-1}, \nu_k)$ *alone*. Consider a time t, and choose l such that $t \in [\nu_l, \nu_{l+1})$. Using (4.1) and the above upper bounding technique, we obtain

$$\begin{aligned} Q_i(t) &= A_i(t) - \int_0^t r_i(z)\mathbf{I}(Q_i(z) > 0)dz \\ &\leq \sum_{k=0}^{l-2} \left[A_i(\nu_{k+1}) - A_i(\nu_k) - \int_{\nu_{k+1}}^{\nu_{k+2}} r_i(z)dz \right]_+ \\ &\quad + A_i(t) - A_i(\nu_{l-1}), \end{aligned} \tag{4.16}$$

where $[\theta]_+ = \max(0, \theta)$.

For each interval $[\nu_k, \nu_{k+1})$, define the following two random variables:

$$\alpha_i(k) = \frac{A_i(\nu_{k+1}) - A_i(\nu_k)}{\Gamma_k}, \text{ and}$$

$$\beta_i(k) = \frac{1}{\Gamma_k} \int_{\nu_k}^{\nu_{k+1}} r_i(z)dz.$$

It follows from the strong law of large numbers that, with probability 1, $\lim_{k\to\infty} \alpha_i(k) = \lambda_i$. From Lemma 4.4 and ergodic theorem for Markov chains, it follows that, with probability 1, $\lim_{k\to\infty} \beta_i(k+1) = \lambda_i$. Since the arrival process $A_i(t)$ is non-decreasing and the increments are bounded by K, we have

$$\begin{aligned} A_i(t) - A_i(\nu_{l-1}) &\leq A_i(\nu_{l+1}) - A_i(\nu_{l-1}) \\ &\leq K(\nu_{l+1} - \nu_{l-1}) \\ &= K(\Gamma_{l-1} + \Gamma_l). \end{aligned} \tag{4.17}$$

Rewriting (4.16) with above defined random variables and applying (4.17) along with $\nu_l \leq t$ and $\Gamma_k \leq \Gamma_{k+1}$, we obtain

$$\frac{Q_i(t)}{t} \leq \frac{1}{\nu_l} \sum_{k=0}^{l-2} \Gamma_k \left[\alpha_i(k) - \beta_i(k+1)\right]_+ + \frac{K(\Gamma_{l-1} + \Gamma_l)}{\nu_l}. \tag{4.18}$$

In (4.18), the second term on the right hand side (RHS) goes to zero as $\Gamma_l/\nu_l \to 0$ as $l \to \infty$. The first term on the RHS of (4.18) goes to zero with probability 1 as $\alpha_i(k) - \beta_i(k+1) \to 0$, $\nu_l \geq \sum_{k=0}^{l-2} \Gamma_k$ and $\nu_l \to \infty$. Thus, for any given $\boldsymbol{\lambda} \in \mathcal{R}_c^o$, with probability 1,

$$\lim_{t\to\infty} \frac{Q_i(t)}{t} = 0, \forall i \in \mathcal{L},$$

which completes the proof. □

This result is important due to the following two reasons.

1. The result shows that this algorithm has good performance, and an algorithm that approaches the operating point of this algorithm has the

potential to perform "well." Essentially, this aspect is utilized to obtain the adaptive algorithm.

2. The non-adaptive algorithm does not require the knowledge of the number of nodes or ϵ, as required by the adaptive algorithm. This suggests the existence of similar gradient-like algorithms that perform "well" with different algorithm parameters that may not depend on the number of nodes or ϵ. We do not explore this further, but the non-adaptive algorithm will serve as the starting point to address such issues.

4.6 Throughput Optimality of the Adaptive Algorithm

In this section, we establish the throughput-optimality of the adaptive algorithm for a particular choice of parameters. The algorithm parameters used in this section are dependent on the number of links n and ϵ. It is evident from the theorem that ϵ determines how close the algorithm is to optimal performance. Define

$$C(n) = 3^5(2\bar{K} + K)^2 \left(\frac{\bar{K}^2 n^2}{2} + n \right).$$

We set all the step sizes (irrespective of interval) to

$$\alpha_l = \alpha(n, \epsilon) = \epsilon^2 / C(n), \tag{4.19}$$

and D used in the projection to

$$D = D(n, \epsilon) = \frac{16\bar{K}}{\underline{K}} \frac{n}{\epsilon} \log \left\lceil \frac{2\bar{K}}{\epsilon} \right\rceil + \bar{K}. \tag{4.20}$$

All the interval lengths (irrespective of interval) are set to

$$T_l = T(n, \epsilon) = \exp\left(\hat{K}\left(\frac{n^2}{\epsilon}\log\frac{n}{\epsilon}\right)\right) \tag{4.21}$$

for some large enough constant $\hat{K} > 0$.

Remark 4.2. *The large value of $T(n, \epsilon)$ in (4.21) is due to the poor bound on the conductance of the rate allocation Markov chain. The parameters given by (4.19), (4.20) and (4.21) are one possible choice of the parameters. We would like to emphasize that this choice is primarily for the purpose of the proofs. The choice of right parameters (and even the update functions) in practice are subject to further study especially based on the network configuration and delay requirements. Some comments on this are given in Section 4.8.*

We start with the optimization framework developed in the previous section. For the adaptive algorithm, the relevant optimization problem is as follows: given $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda} + \frac{\epsilon}{2}\mathbf{1} \in \mathcal{R}_c$,

$$\begin{aligned} &\text{maximize} \quad F_\epsilon(\mathbf{v}) := F\left(\mathbf{v}, \boldsymbol{\lambda} + \frac{\epsilon}{4}\mathbf{1}\right) \\ &\text{subject to} \quad \mathbf{v} \in \mathbb{R}^n. \end{aligned} \tag{4.22}$$

The following result is an extension of Lemma 4.4.

Lemma 4.5. *Consider any given $\epsilon > 0$ and $\boldsymbol{\lambda}$. Then, the optimization problem in (4.22) is strictly concave in $\mathbf{v}$ with gradient $\nabla F_\epsilon(\mathbf{v}) = \boldsymbol{\lambda} + \frac{\epsilon}{4}\mathbf{1} - \mathbf{s}_\mathbf{v}$ and Hessian*

$$H(F(\mathbf{v})) = -\left(\mathbb{E}_{\pi_\mathbf{v}}\left[\mathbf{r}\mathbf{r}^T\right] - \mathbb{E}_{\pi_\mathbf{v}}\left[\mathbf{r}\right]\mathbb{E}_{\pi_\mathbf{v}}\left[\mathbf{r}^T\right]\right).$$

Further, let $\boldsymbol{\lambda} + \frac{\epsilon}{2}\mathbf{1} \in \mathcal{R}_c$. *Then, it has a unique solution* $\mathbf{v}^*$, *which is finite, such that the offered service rate vector under* $\mathbf{v}^*$ *is equal to* $\boldsymbol{\lambda} + \frac{\epsilon}{4}\mathbf{1}$, *i.e.,* $\mathbf{s}_{\mathbf{v}^*} = \boldsymbol{\lambda} + \frac{\epsilon}{4}\mathbf{1}$. *In addition, if* $\epsilon \leq 4\lambda_{\min}$, *then the optimal value* $\mathbf{v}^*$ *is such that*

$$\|\mathbf{v}^*\|_\infty \leq \frac{16\bar{K}}{\underline{K}} \frac{n}{\epsilon} \log \left\lceil \frac{2\bar{K}}{\epsilon} \right\rceil . \tag{4.23}$$

Proof. Please see Appendix B.2. □

The update step in (4.9), which is central to the adaptive algorithm, can be intuitively thought of as a gradient decent technique to solve the above optimization problem. Technically, it is different as the arrival rate and offered service rate are replaced with their empirical values for decentralized operation. The algorithm parameters can be chosen in order to account for this. This forms the central theme of this section.

4.6.1 Mixing within update interval

Consider a time interval $[\tau_l, \tau_{l+1})$. During this interval the algorithm uses parameters $v_i(\tau_l)$. For simplicity, in this subsection, we denote $v_i(\tau_l)$ by v_i and the vector by $\mathbf{v}$ and $\mathbb{E}[\cdot|\mathbf{v}]$ by $\mathbb{E}[\cdot]$. For the rate allocation Markov chain (MC) introduced in Section 4.5, we obtain an upper bound on the convergence time or the mixing time.

To obtain this bound, we perform *uniformization* of the continuous-time MC (CTMC) and use results given in Section 4.3 on the mixing time of discrete-time MC (DTMC). The uniformization constant used is

$$A = n \exp(\bar{K}\|\mathbf{v}\|_\infty).$$

The resulting DTMC has the same state-space $\mathcal{R}$ with transition probability matrix P. The transition probability from state $\hat{\mathbf{r}} \in \mathcal{R}$ to state $\mathbf{r} \in \mathcal{R}, \mathbf{r} \neq \hat{\mathbf{r}}$ is $P(\hat{\mathbf{r}}, \mathbf{r}) = q(\hat{\mathbf{r}}, \mathbf{r})/A$, and from state $\hat{\mathbf{r}} \in \mathcal{R}$ to itself is $P(\hat{\mathbf{r}}, \hat{\mathbf{r}}) = 1 + q(\hat{\mathbf{r}}, \hat{\mathbf{r}})/A$. With our choice of parameters $u_{i,j}$ given by (4.12), we can simplify (4.10) to

$$f(\hat{\mathbf{r}}, \mathbf{r}) = \exp\left(\sum_{i=1}^{n} r_i v_i \mathbf{I}(r_i \neq \hat{r}_i)\right). \tag{4.24}$$

For all $\hat{\mathbf{r}}, \mathbf{r} \in \mathcal{R}, \mathbf{r} \neq \hat{\mathbf{r}}$, clearly $q(\hat{\mathbf{r}}, \mathbf{r}) \leq \exp(\bar{K}\|\mathbf{v}\|_\infty)$. Since at most n elements in every row of the transition rate matrix of the CTMC is positive $|q(\hat{\mathbf{r}}, \hat{\mathbf{r}})| \leq A$ for all $\hat{\mathbf{r}} \in \mathcal{R}$. Therefore, P is a valid probability transition matrix.

The DTMC has the same stationary distribution as the CTMC. In addition, the CTMC and the DTMC have one-to-one correspondence through an underlying independent Poisson process with rate A. In this subsection, time t denotes the time within the update interval, i.e., $t = 0$ denotes global time τ_l. Let $\boldsymbol{\mu}(t)$ be the distribution over $\mathcal{R}$ given by the CTMC at time t, and ζ be a Poisson random variable with parameter At. Then, we have

$$\begin{aligned} \boldsymbol{\mu}^T(t) &= \sum_{m \in \mathbb{Z}_+} \Pr(\zeta = m) \boldsymbol{\mu}^T(0) P^m \\ &= \boldsymbol{\mu}^T(0) \exp(At(P - I)), \end{aligned} \tag{4.25}$$

where I is the identity matrix. Next, we provide the upper bound on the mixing time of the CTMC.

Lemma 4.6. *Consider any $\rho_1 > 0$. Then, there exists a constant $K_1 > 0$, such that, if*

$$t \geq \exp\left(K_1\left(n\|\mathbf{v}\|_\infty + n\log\frac{1}{\epsilon}\right)\right)\log\frac{1}{\rho_1}, \tag{4.26}$$

then the total variation between the probability distribution $\boldsymbol{\mu}(t)$ at time t given by (4.25) and the stationary distribution $\boldsymbol{\pi}_{\mathbf{v}}$ given by (4.14) is smaller than ρ_1, i.e., $\|\boldsymbol{\mu}(t) - \boldsymbol{\pi}_{\mathbf{v}}\|_{TV} \leq \rho_1$.

Proof. Please see Appendix B.3. □

Lemma 4.6 is used to show that the error associated with using empirical values for arrival rate and offered service rate in the update rule (4.9) can be made arbitrarily small by choosing large enough T. This is formally stated in the next lemma.

Lemma 4.7. *Consider $\rho_2 > 0$. Then, there exists a constant $K_2 > 0$, such that, if the updating period*

$$T \geq \exp\left(K_2\left(n\|\mathbf{v}\|_\infty + n\log\frac{1}{\epsilon}\right)\right)\frac{1}{\rho_2},$$

then for any time interval $[\tau_l, \tau_{l+1})$

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{\lambda}}(l) - \boldsymbol{\lambda}\right\|_1\right] + \mathbb{E}\left[\|\hat{\mathbf{s}}(l) - \mathbf{s}_{\mathbf{v}}\|_1\right] \leq \rho_2. \tag{4.27}$$

Proof. Please see Appendix B.4. □

Thus, the important result is that due to the mixing of the rate allocation Markov chain, the empirical offered service rate is *close* to the offered service rate. The next step is to address whether the offered service rates over multiple update intervals is *higher* than the arrival rates.

4.6.2 'Drift' over multiple update intervals

We consider multiple update intervals, and establish that the average empirical offered service rate is *strictly* higher than the arrival rate. This result follows from the observation that, if the error in approximating the true values by empirical values are sufficiently small, then the average of the expected value of the gradient of $F_\epsilon(\mathbf{v})$ over sufficiently large number of intervals should be small. In this case, we can expect the average offered service rate to be close to $\mathbf{s}_{\mathbf{v}^*}$. Since, $\mathbf{s}_{\mathbf{v}^*}$ is strictly higher than arrival rates, we can expect the average offered service rate to be strictly higher than the arrival rate. The result is formally stated next.

Lemma 4.8. *Consider $N(n,\epsilon) = (7 \times 3^5 n D^2)/(\alpha\epsilon^2)$ update intervals. Then, the average of empirical service rates over these update intervals is greater than or equal to $\boldsymbol{\lambda} + \frac{\epsilon}{8}\mathbf{1}$, i.e.,*

$$\frac{1}{N}\sum_{l=1}^{N} \mathbb{E}\left[\hat{\mathbf{s}}(l)\right] \geq \boldsymbol{\lambda} + \frac{\epsilon}{8}\mathbf{1}.$$

Proof. Please see Appendix B.5. □

Now, we proceed to show that the appropriate 'drift' required for stability is obtained.

4.6.3 Proof of Theorem 4.2

Consider the underlying network Markov chain $X(l)$ consisting of all the queues in the network, the update parameters, and the resulting rate

allocation vectors at time τ_l, i.e., $X(l) = (\mathbf{Q}(\tau_l), \mathbf{v}(\tau_l), \mathbf{r}(\tau_l))$ for $l \in \mathbb{Z}_+$. It follows from the system model and the algorithm description that $X(l)$ is a time-homogenous Markov chain on an uncountable state-space $\mathcal{X}$. The σ-field on $\mathcal{X}$ considered is the Borel σ-field associated with the product topology. For more details on dealing with general state-space Markov chains, we refer readers to [86].

We consider a Lyapunov function $V : \mathcal{X} \to \mathbb{R}_+$ of the form, $V(\mathbf{x}) = \sum_{i=1}^{n}(Q_i^2 + v_i^2 + r_i^2)$ for $\mathbf{x} = (\mathbf{Q}, \mathbf{v}, \mathbf{r})$. In order to establish positive Harris recurrence, for any $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda} + \epsilon\mathbf{1} \in \mathcal{T}$, we use multi-step[9] Lyapunov and Foster's *drift* criteria to establish positive recurrence of a set of the form $V(\mathbf{x}) \leq \kappa$, for some $\kappa > 0$. From the assumption on the arrival processes, it follows that $V(\mathbf{x}) \leq \kappa$ is a closed *petite* (small) set (for definition and details see [57, 86]). Intuitively, petite set is a generalization of a recurrent state for a countable Markov chain to an uncountable Markov chain. It is well known that these two results imply positive Harris recurrence [86].

Next, we obtain the required drift criteria. For simplicity, we denote $\mathbb{E}[\cdot|X(0)]$ by $\mathbb{E}[\cdot]$ in the rest of this section. Consider

$$\begin{aligned}
\mathbb{E}\left[Q_i^2(TN) - Q_i^2(0)\right] &= \mathbb{E}\left[(Q_i(TN) - Q_i(0))^2\right. \\
&\quad \left. +2Q_i(0)(Q_i(TN) - Q_i(0))\right] \\
&\overset{(a)}{\leq} \left(\max\left(K, \bar{K}\right)TN\right)^2 + \\
&\quad 2Q_i(0)\,\mathbb{E}\left[Q_i(TN) - Q_i(0)\right].
\end{aligned}$$

[9]This is a special case of the state-dependent drift criteria in [86].

Here, (a) follows from the fact that over unit time queue difference belong to $[-\bar{K}, K]$. Now, we look at two cases. If $Q_i(0) > \bar{K}TN$, clearly $Q_i(t) > 0$ during interval $[0, TN]$ as service rate is less than or equal to $\bar{K}$. For this case, from Lemma 4.8,

$$\begin{aligned} 2Q_i(0)\,\mathbb{E}\left[Q_i(TN) - Q_i(0)\right] &= 2Q_i(0)\,T\left(\sum_{l=1}^{N}\left(\lambda_i - \mathbb{E}\left[\hat{s}_i(l)\right]\right)\right) \\ &\leq -\frac{\epsilon}{4}TNQ_i(0) \\ &\overset{(a)}{\leq} -\frac{\epsilon}{4}TNQ_i(0) + \frac{\epsilon}{4}\bar{K}(TN)^2. \end{aligned}$$

Here, (a) is trivial, but the extra term is added to ensure that the RHS evaluates to a non-negative value for $Q_i(0) \leq \bar{K}TN$. If $Q_i(0) \leq \bar{K}TN$, then clearly

$$2Q_i(0)\mathbb{E}[Q_i(TN) - Q_i(0)] \leq 2\bar{K}K(TN)^2.$$

Since the bounds for each case do not evaluate to negative values for the other case, we have

$$\mathbb{E}\left[Q_i^2(TN) - Q_i^2(0)\right] \leq -\frac{\epsilon}{4}TNQ_i(0) + \left(\left(K + \bar{K}\right)^2 + \frac{\epsilon}{4}\bar{K}\right)(TN)^2.$$

Since both $\mathbf{v}$ and $\mathbf{r}$ are bounded, there exists some fixed $M(n, \epsilon)$ such that

$$\mathbb{E}\left[v_i^2(TN) - v_i^2(0)\right] + \mathbb{E}\left[r_i^2(TN) - r_i^2(0)\right] \leq M(n, \epsilon).$$

Summing up over all $i \in \mathcal{L}$, we obtain

$$\begin{aligned} \mathbb{E}\left[V(X(N)) - V(X(0))\right] \leq\ & -\frac{\epsilon}{4}TN\left(\sum_{i=1}^{n}Q_i(0)\right) \\ & + nM(n,\epsilon) + n\left((K + \bar{K})^2 + \frac{\epsilon}{4}\bar{K}\right)(TN)^2. \end{aligned}$$

This shows that there exists some $\kappa > 0$ such that for all $\mathbf{x}$ with $V(\mathbf{x}) > \kappa$ there is strict negative drift. Hence, the set $V(\mathbf{x}) \leq \kappa$ is positive recurrent. Since $\boldsymbol{\lambda} + \frac{\epsilon}{2}\mathbf{1} \in \mathcal{R}_c$, clearly $\boldsymbol{\lambda} + \epsilon\mathbf{1} \in \mathcal{T}$. This completes the proof. □

In summary, given any rate region for a wireless network, the (queue-length based) algorithm has $(1 - \epsilon)$-optimal performance.

4.7 Application: White-Space Networks

An important application of our algorithmic framework is in the domain of white-space networks [42, 87]. White-space radios are typically required to sense the environment [3]. Therefore, these radios are designed with highly accurate sensing capabilities. Even though these are primarily designed for sensing the presence of primary radios, the same capability can be exploited for sensing secondary radios. In this section, we consider a network of secondary nodes that use the same spectrum, but different from that used by primary nodes. In particular, we assume that the secondary nodes have already found spectrum that are not utilized by primary nodes.

Since such a white-space network of secondary nodes is not centrally controlled, it is desirable to obtain simple distributed algorithms. However, the scheduling problem in these white-space networks is different from the link scheduling problem in traditional wireless networks [34]. First, the available spectrum for the operation of this network is fragmented with different propagation characteristics. Second, these secondary nodes are usually equipped with multiple radios to operate simultaneously in different bands. This is re-

ferred to as the multiple-band multiple-radio scheduling problem. Next, we describe the multiple-band multiple-radio scheduling problem in detail.

Consider the network model introduced in Section 4.2. Define functions $s : \mathcal{L} \mapsto \mathcal{N}$ that maps links to source nodes, and $d : \mathcal{L} \mapsto \mathcal{N}$ that maps links to destination nodes. The available spectrum for the operation of this network is fragmented. The spectrum consists of M bands, labeled $\mathcal{B} = \{1, 2, \ldots, M\}$, with bandwidths $B_1, B_2, \ldots, B_M$. The transmission from a node to another node gets different spectral efficiencies on different bands. For a link i, let $c_{i,b}$ be the spectral efficiency that node $s(i)$ gets when it transmits on band b to node $d(i)$. The link interference graphs are also different on different bands. Let $\mathcal{G}_b = (\mathcal{L}, \mathcal{E}_b)$ be the link interference graph on band b, i.e, the transmission of link u interfere with the transmission of link v in band b if $(u, v) \in \mathcal{E}_b$. We assume that the link interference is symmetric, i.e., if $(u, v) \in \mathcal{E}_b$ then $(v, u) \in \mathcal{E}_b$. These capture the frequency dependent propagation characteristics and the spatial variation of the quality of spectrum. Further, each node j is equipped with a_j radios.

At time t, the decision whether link i is operated in band b is represented by binary decision variables $\sigma_{i,b}(t)$, with 1 representing "true" and 0 representing "false". The decision variables have to satisfy the constraints that arise from the following. (i) Interference constraints: In every band, the set of allocated links must be non-interfering. (ii) Radio constraints: The total number of radios at each node is limited, and these radios are half-duplex, i.e., a link requires its end nodes to dedicate one radio each for a transmission to

happen. More formally, the set of constraints are:

$$\begin{aligned} \sigma_{u,b}(t) + \sigma_{v,b}(t) &\leq 1, \forall (u,v) \in \mathcal{E}_b, \forall b \in \mathcal{B}, \\ \sum_{i:j \in \{s(i),d(i)\}} \sum_{b \in \mathcal{B}} \sigma_{i,b}(t) &\leq a_j, \forall j \in \mathcal{N}. \end{aligned}$$

For a feasible schedule, the rate of flow supported on link i is

$$r_i(t) = \sum_{b \in \mathcal{B}} \sigma_{i,b}(t) c_{i,b} B_b.$$

We denote the vector of above rates by $\mathbf{r}(t)$. The throughput region $\mathcal{T} \subseteq \mathbb{R}_+^n$ is defined as the convex hull of the set of all feasible rate vectors. Note that the queue dynamics is exactly same as described in Section 4.2.

4.7.1 Distributed Algorithm

In this section, we present an adaptation of the developed algorithm that is throughput-optimal for multiple-band multiple-radio scheduling. For simplicity, we assume that perfect and instantaneous carrier sensing is possible on every band. The scheduling vector corresponding to link i is $\boldsymbol{\sigma}_i(t) = \{\sigma_{i,b}(t)\}_{b \in \mathcal{B}}$. For this link, the possible states are

$$\{\boldsymbol{\theta}_i : \boldsymbol{\theta}_i = \{\theta_{i,b}\}_{b \in \mathcal{B}}, \theta_{i,b} \in \{0,1\}, \|\boldsymbol{\theta}_i\|_0 \leq \min\{a_{s(i)}, a_{d(i)}\}\}.$$

The link uses an independent exponential clock corresponding to each state with transition rate $\exp(\sum_{b \in \mathcal{B}} \theta_b c_{i,b} B_b v_i)$ for state $\boldsymbol{\theta}$. Based on these clocks, the link obtains $\boldsymbol{\sigma}_i(t)$ as follows:

1. If the clock associated with a state (say $\boldsymbol{\theta}$) ticks and transitioning to that state $\boldsymbol{\sigma}_i(t) = \boldsymbol{\theta}$ is feasible,[10] then $\boldsymbol{\sigma}_i(t)$ is changed to $\boldsymbol{\theta}$;

2. Otherwise, $\boldsymbol{\sigma}_i(t)$ remains the same.

The above procedure continues. The parameter v_i is updated over time as a function of the queue-length $Q_i(t)$ as described in Section 4.4. This makes the algorithm completely distributed. The vector of $\{v_i\}_{i\in\mathcal{L}}$ is denoted by $\mathbf{v}$.

In order to establish that this algorithm is throughput-optimal, we show a correspondence between it and the rate allocation algorithm in Section 4.4. Consider a fixed $\mathbf{v}$. The above algorithm forms a Markov chain on the set of feasible states. Let $S(t)$ denote the matrix formed by vectors $\{\boldsymbol{\sigma}_i(t)\}_{i\in\mathcal{L}}$, and $\mathcal{S}$ denote the set of feasible matrices satisfying (4.28) and (4.28). The transition rate from state $\hat{S} = \{\hat{\boldsymbol{\theta}}_i\}_{i\in\mathcal{L}}$ to state $S = \{\boldsymbol{\theta}_i\}_{i\in\mathcal{L}}$ can be expressed as

$$q(\hat{S}, S) = \begin{cases} f(\hat{S}, S), & \text{if } \sum_{i=1}^{n} \mathbf{I}(\boldsymbol{\theta}_l \neq \hat{\boldsymbol{\theta}}_i) = 1, \\ 0, & \text{if } \sum_{i=1}^{n} \mathbf{I}(\boldsymbol{\theta}_i \neq \hat{\boldsymbol{\theta}}_i) > 1, \end{cases}$$

where

$$f(\hat{S}, S) = \exp\left(\sum_{i=1}^{n}\sum_{b\in\mathcal{B}} \theta_{i,b} c_{i,b} B_b v_i \mathbf{I}(\boldsymbol{\theta}_i \neq \hat{\boldsymbol{\theta}}_i)\right).$$

And, the diagonal elements of the rate matrix are given by

$$q(\hat{S}, \hat{S}) = -\sum_{S\in\mathcal{S}, S\neq\hat{S}} q(\hat{S}, S), \forall \hat{S} \in \mathcal{S}.$$

Now, the following lemma is immediate.

[10] This is determined using carrier sensing and radio constraints at the source and the destination of that link.

Lemma 4.9. *The Markov chain* $(\mathcal{S}, q)$ *is reversible and has the stationary distribution*

$$\begin{aligned}\pi_{\mathbf{v}}(S) &= \frac{\exp\left(\sum_{i=1}^{n}\sum_{b\in\mathcal{B}}\theta_{i,b}c_{i,b}B_b v_i\right)}{\sum_{\tilde{S}\in\mathcal{S}}\exp\left(\sum_{i=1}^{n}\sum_{b\in\mathcal{B}}\hat{\theta}_{i,b}c_{i,b}B_b v_i\right)} \\ &= \frac{\exp\left(\mathbf{r}(S)\cdot\mathbf{v}\right)}{\sum_{\tilde{S}\in\mathcal{S}}\exp\left(\mathbf{r}(\hat{S})\cdot\mathbf{v}\right)}.\end{aligned}$$

Furthermore, this Markov chain converges to this stationary distribution starting from any initial distribution.

The *offered service rate* vector under the stationary distribution is $\mathbf{s_v} = \sum_{S\in\mathcal{S}}\pi_{\mathbf{v}}(S)\mathbf{r}(S)$. Thus, we show a one-to-one correspondence to the rate allocation algorithm. As a consequence, we establish the throughput-optimality of the algorithm described in this section based on Theorem 4.2.

4.8 Further Discussion & Simulation

4.8.1 Convex Rate Regions

Definition 4.8 (Low delay)**.** *We say that a rate allocation algorithm has low-delay if, for any given* $\epsilon > 0$*, the average queue-lengths are polynomial in number of links* n *and* $\log 1/\epsilon$ *for all* $\boldsymbol{\lambda}$ *such that* $\boldsymbol{\lambda} + \epsilon\mathbf{1} \in \mathcal{T}$*.*

For arbitrary rate regions, there are negative results based on computational hardness for the existence of a low-delay throughput-optimal algorithm [109]. Intuitively, a rate allocation algorithm needs to operate at multiple operating points leading to potentially large-delay for switching from one operating point to another.

However, for convex rate regions, these hardness results do not hold. It is in fact straightforward to obtain an algorithm that is throughput-optimal and has low-delay in this setting. The algorithm simply needs to adapt to a single operating point that is "close" to the arrival-rate vector. Hence, we can estimate the arrival rates and operate based on this single point. This observation emphasizes the fact that the hardness lies in convexifying an arbitrary non-convex rate region in a decentralized manner.

4.8.2 Determining Rate Feasibility

Although the rate allocation algorithm removes the control overhead associated with queue-length exchanges in the network, it still requires each link to determine rate feasibility. To elaborate, feasibility implies data-rates of other links are not impacted, i.e., other links are able to maintain their data-rates in spite of the change in the given link's data-rate. Each link can possibly change its coding and modulation strategies to ensure this. This requires design of dynamic coding and modulation to support a constant rate. A link can determine whether a data-rate is feasible if it knows the current set of data-rates associated with other links. An important fact that makes the algorithm of practical value is that a link needs to know only data-rates associated with those links that it "interferes with". Even though all links interfere with each other in a wireless network, the magnitude of interference decays with distance. Therefore, in a large network, every link needs to learn data-rates associated with few physically near-by links from control messages, for example, through

ACK/NACKs when ARQ is present. Clearly, this is an approximation, but this is a widely used approximation while designing algorithms for large networks. We refer to the process of determining rate feasibility from the interactions of physically near-by links as *channel measuring*. This can be considered as a natural extension of *sensing* in CSMA.

In order to further explain *channel measuring*, we consider an example with a simplified physical-layer model. In this model, a transmitter can potentially communicate with a receiver if the receiver is within distance d_0. This transmitter can communicate at data-rate $r_j, 1 \leq j \leq k$, if there are no other transmitters within distance d_j to it. We consider $r_1 \leq r_2 \leq \ldots \leq r_k$ and $d_0 \leq d_1 \leq \ldots \leq d_k$. In this setting, for channel measuring, a transmitter needs to simply determine the distance of the nearest active transmitter. Even though we used an over simplified physical-layer model, this shows that channel measuring is a very natural technique for determining rate feasibility. Furthermore, it suggests that slightly more complicated schemes than carrier sensing may be enough to obtain significant throughput gains.

For complex physical-layer interactions, we acknowledge that channel measuring requires well-designed physical-layer control architecture, which, by itself, is a fairly non-trivial problem. However, radios that perform complex physical-layer signaling are increasingly common and each node has access to current channel interference level, information from beacons, pilot signals and its own location. These will definitely help such radios to perform channel measuring using existing physical-layer control overhead.

In the context of 802.11 networks, we have actually implemented and tested the algorithmic framework developed so far. By using a combination of state measurement techniques, we show in [11] that queue and channel based rate allocation is indeed practically feasible. Thus, it validates our theoretical framework in the context of WiFi networks.

4.8.3 Simulation

So far, we show that the algorithm provides throughput-optimal performance for particular choice of algorithm parameters. Although this has significant theoretical value, these parameters may not be directly suitable in practice. In particular, we may have to limit the update interval length and attempt rates as large values of update interval can result in large queue-lengths, and large attempt rates can result in frequent changes in data-rates. There is certain hardware and physical-layer coding limitations on frequently changing data-rates, and frequent attempts lead to increased sensing/measuring overhead. These limitations can be easily dealt with through modified algorithm parameters.

Our approach motivates a more general class of algorithms that can be throughput-optimal for appropriate choice of parameters. We can consider the general class with update rule

$$v_i(\tau_{l+1}) = h\left(v_i(\tau_l), \hat{\lambda}_i(l) - \hat{s}_i(l)\right)$$

for some function $h(\cdot)$. Next, we provide a "good" choice of this function based on simulation results. We consider the update rule $v_i(\tau_{l+1}) = \log(1 + Q(\tau_{l+1}))$.

The log function results in linear update near origin and prevents the rapid growth of v_i with queue-length. This update rule has been used in literature [91]. Thus, the purpose of this section is not to simulate the adaptive update rule that is provably throughput-optimal.

Consider the same Gaussian multiple access channel example with two links as before. This is shown in Figure 4.1. This is simply an illustrative example to show scheduling over multiple data-rate levels. Similar simulation results apply for any number of users. Let the average power constraint at the transmitters be $P = 3$ and noise variance at the receiver be $N = 1$. The information-theoretic capacity region of this channel is the pentagon shown in Figure 4.2 where $C(x) = 0.5 \log_2(1 + x)$. The set of rate levels chosen by both transmitters are $\{0, a, b\}$ where $a = 0.4$ and $b = 1$. The only infeasible rate allocation pair is $[1, 1]$. Consider the following arrival processes at both the transmitters. At integral times, the queues are incremented by an i.i.d. Bernoulli random variable such that the arrival rate is $\lambda = \rho\frac{a+b}{2}$, where $\rho > 0$ represents the load in the system. Clearly, the network will be unstable for $\rho > 1$.

For this system, we perform Monte-Carlo simulations with update interval $T = 10$. We provide a trace of the queue-length process for $\rho = 0.9$ and $\rho = 1.1$ in Figure 4.4. We observe that the algorithm supports 90% load in the system without large increase in queue-lengths. Intuitively, this symmetric operating point is one of the difficult operating points for a distributed algorithm. More importantly, the sum-rate $\rho(a + b)$ obtained is close to the

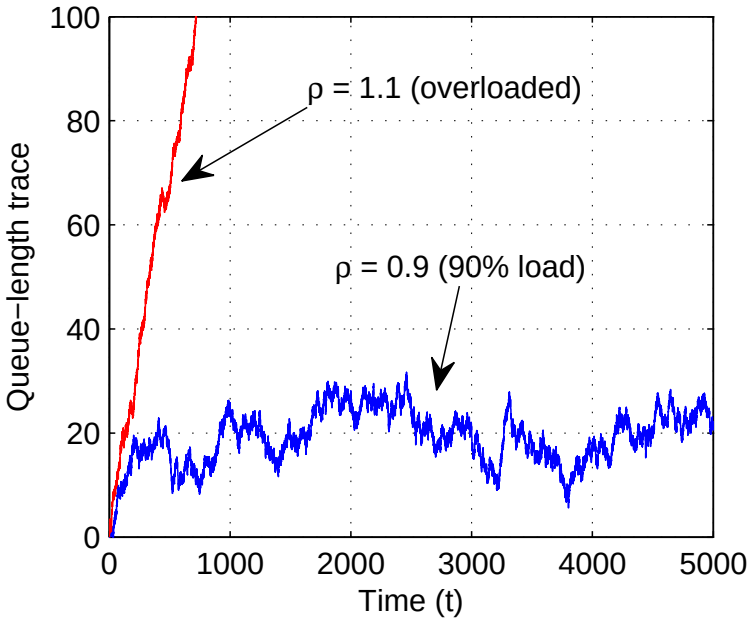

Figure 4.4: Queue-length trace from simulation

information-theoretic sum-capacity of this system.

Chapter 5

Genie-MAC Outer Bound for Gaussian Interference Networks

5.1 Introduction

The motivation to understand the fundamental limits of interference channel arises from its relevance in wide range of applications. However, large gaps exist in our understanding of interference channels (IFCs). Since the introduction of interference channels [4], the class of two-transmitter two-receiver interference channels have been studied in great detail. Initial successes were achieved in the domains of the two-user very strong [22] and strong discrete-memoryless [29], and Gaussian IFCs [103]. These results found that the capacity of the two-user strong Gaussian IFCs is achieved as the intersection of the capacity region of two multiple access channels. In brief, it is optimal for each receiver to attempt to *jointly* decode both messages. Generalizing the notion of strong interference and such a capacity result to more than two users is not straightforward. Moreover, joint decoding of messages can be easily shown to be not optimal for weak (i.e. not-strong) interference channels. In general, for weak Gaussian IFCs, intuition suggests that treating at least a portion of the interference as noise is advisable. Indeed, recent results have shown that, for very weak Gaussian interference channels [7, 88, 112], treating

the entire interference as noise is optimal from a sum-capacity perspective. The well-known Han-Kobayashi region [51] combines elements of both, by enabling each receiver to perform joint-decoding of a portion of the interference with the signal. To date, it remains the best-known achievable region for the two-user interference channel.

Deriving exact capacity results for any channel requires good outer bounds. To this end, genie-aided bounds have played a central role in interference channel literature, with a majority of capacity results known resulting from it. The sum-capacity of two-user very-weak Gaussian IFCs as established in [7, 88, 112] was shown using genie-aided techniques. Similarly, the capacity region of the Gaussian interference channel has been characterized to within one bit using Gaussian codebooks and the Han-Kobayashi region for the inner bound and a genie-aided technique for the outer bound in [41]. As a concept, the idea of genie-aided outer bounds is fairly intuitive. However, the main challenge lies in determining the right genie and optimizing the resulting outer bound to show that it matches (or comes close to matching) the desired achievable rates. Finding the right genie and matching rate expressions is non-trivial, which forms an important part of the work in [7, 41, 76, 88, 112]. One of the contributions is similar - where we obtain an optimization problem formulation for genie-aided bounds for a Gaussian K-user IFC that is analytically tractable.

A majority of known capacity results in the domain of interference channels are for two-user cases. Unlike the Gaussian multiple access channel

(MAC) and the Gaussian broadcast channel (BC), the extension from two-user to $K > 2$ in Gaussian IFCs is not straightforward. The limits of K-user IFCs has received considerable attention in recent years. In developing achievable schemes, the concept of *interference alignment* has been developed and applied to numerous channel settings. As introduced in [18], interference alignment minimizes the subspace spanned by the net interference seen at each receiver, thus achieving $K/2$ total degrees of freedom (DoF) for the system. This is interesting in that linear growth in capacity is achieved asymptotically in signal to noise ratio (SNR) and can be shown to be optimal. Although the initial results on DoF analysis for interference channels was for time-varying channels, more recently results have shown that, for constant channels as well, under certain conditions, a DoF of $K/2$ can be achieved [89]. However, a majority of existing literature is on the DoF of K user IFCs, i.e., is asymptotic in SNR, with only a limited set of results known for finite SNR.

Recently, some initial results have been developed for finite SNR. To this end, ergodic interference alignment [90] and lattice interference alignment [56] have been developed for time varying and static channel settings respectively. Ergodic alignment, under certain conditions on channel symmetry over time, achieves a sum-rate of $K/2\log(1+2SNR)$ non-asymptotically. Similarly, lattice alignment is found to achieve a sum-rate of $D\log(SNR)$ where D is a channel-dependent parameter that grows with K for a number of static channel settings. However, each of these results makes certain assumptions on the channel. Ergodic alignment requires time variations and symmetry, while

lattice alignment assumes the channel to be "well conditioned". Moreover, neither scheme is known to be sum-rate optimal for the Gaussian K-user IFC.

5.1.1 Our Approach & Contributions

We develop an outer bound that is both computationally tractable and non-trivial. As discussed above, obtaining such bounds is a fairly non-trivial task. To date, a majority of the bounds for the discrete-memoryless or Gaussian IFC have been developed in the context of two-user IFCs [76]. These two-user bounds can be broadly classified into four types: The broadcast channel type, the multiple-access channel type, the "Z" channel type and the genie-aided type. These four share significant common ground in terms of techniques, and a good understanding of these can be gained from [76].

For more than two users, although each of these bounds can be generalized, they do not necessarily yield non-trivial results. The MAC type bound has been generalized to obtain bounds for general K-user IFCs [19], which are known to be tight (with high probability) in the DoF sense for these channels. However, it is not yet known if they are tight for finite-SNR constant channels. For a class of rational channel gains, combinatorial bounds have been developed that are tighter than the MAC-type bounds in the DoF sense [40]. However, a general theory for developing non-trivial outer bounds for K-user is still missing.

In an attempt to obtain a systematic approach to outer bound for the Gaussian K-user IFC, we build on the literature of genie-aided bounds for

the Gaussian IFC. A genie-aided bound provides receivers in the interference channel with one or more "genies" (side information), thus transforming the channel into one where the rate region can be characterized using single-letter expressions [41, 76]. Genie-aided bounds are fairly general in their structure, and they incorporate central elements of the MAC type and the "Z" interference channel type bounds. Moreover, genie-aided bounds have proven to be effective for characterizing the sum-capacity of very weak interference channels [7, 88, 112]. However, their very general nature is also a disadvantage, as it is not always easy to characterize the bound in closed form.

In view of this, we do the following:

- We develop a class of outer bounds for Gaussian K-user IFCs by analyzing genie-MAC receivers. This class of bounds creates a virtual multiple antenna MAC receiver that outer bounds the capacity region of the original Gaussian IFC.
- We formulate this bound as an optimization problem in terms of channel parameters.
- We apply this bound to the degraded Gaussian IFC. We compare this bound with a simple successive interference cancellation (SIC) based achievable scheme and find that it is sum-capacity optimal.
- We establish an equivalence between the MIMO unit-rank K-user IFC and a corresponding SISO degraded K-user IFC, and thus find the sum-capacity of MIMO unit-rank IFCs.

The class of degraded Gaussian K-user IFCs is fundamentally different from those studied in [18, 89] in that the degrees of freedom of a degraded IFC is one, while those studied in a majority of K-user IFC literature is $K/2$. A degraded Gaussian K-user IFC corresponds to the case where the transfer matrix of the channel is unit-rank. This unit-rank nature may result from a variety of factors, including co-location, usage of a similar medium (digital subscriber line (DSL) or optical) to communicate multiple sources. This is referred to as the *keyhole effect* where the signal from each transmitter passes through the same bottleneck (keyhole) to get to the receiver, thus making the overall channel of low-rank. Unit rank channel matrices represents the other end of the spectrum from full rank matrices - we consider it for two reasons: first, it helps us better understand how dependencies among channel parameters impacts capacity and second, due to analytical tractability - we are able to establish sum-capacity results for this channel for *any* SNR and *any* number of transmit-receive pairs K.

Note that our result includes the previously known result of sum-capacity of degraded two-user Gaussian IFCs [101, 102]. This proof uses a slope-based argument that do not directly generalize to K-user IFCs. It is possible to provide alternate proofs for this result; for example, entropy power inequality (EPI) can be used to arrive at this result. However, our proof is important as it shows that genie-MAC outer bound is sum capacity tight for degraded networks.

5.1.2 Organization

In Section 5.2, we present the system model. In Section 5.3, we characterize an outer bound on the capacity region of any Gaussian K-user IFC. In Section 5.4, we derive the sum-capacity of degraded Gaussian IFCs. In Section 5.5, we extend the results to MIMO unit-rank IFCs.

5.2 System Model

We consider the Gaussian K-user interference channel defined as follows: a communication system consisting of K transmitter-receiver pairs labeled as $1, 2, \ldots, K$. This channel is shown in Figure 5.1. Each transmitter has independent messages intended for the corresponding receiver. At time $t, t \in \mathbb{Z}_+$, the input-output relations that describe the system are:

$$Y_i[t] = \sum_j h_{i,j} X_j[t] + Z_i[t], \quad \forall i. \tag{5.1}$$

Here, $X_j[t]$ is the signal transmitted by the j-th transmitter, $h_{i,j}$ is the constant channel gain from j-th transmitter to i-th receiver, $Z_i[t]$ is the additive white Gaussian noise at i-th receiver, and $Y_i[t]$ is the signal received at the i-th receiver. For simplicity, we consider real valued signal/gain/noise and suppress the time index t henceforth. The power constraint at the j-th transmitter is $\mathbb{E}[X_j^2] \leq P$, and the zero-mean Gaussian noise at all receivers have variance N. The Gaussian K-user IFC is characterized by $\sqrt{P/N}\mathbf{H}$, where $\mathbf{H}$ is the channel matrix with $h_{i,j}$ as the entry corresponding to the i-th row and the j-th column. We use standard information-theoretic definitions for the capacity region and

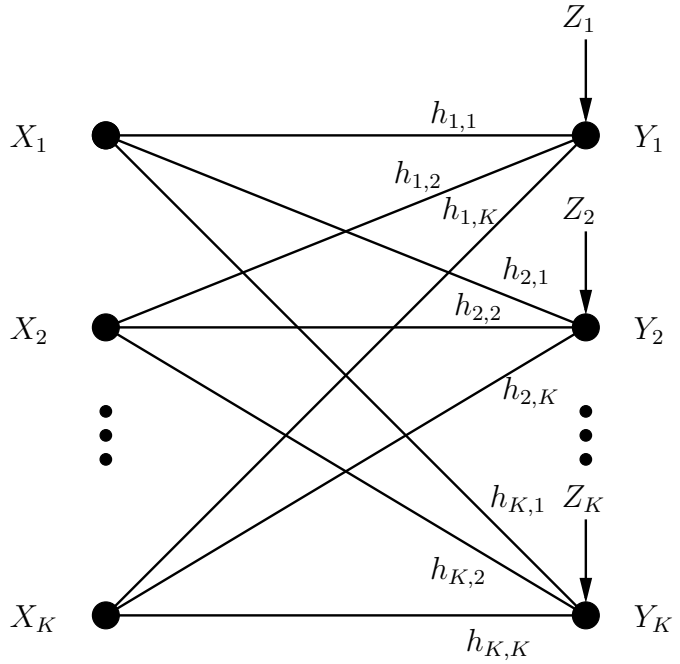

Figure 5.1: Gaussian K-user interference channel

the sum-capacity of this channel. $C^{\mathrm{IC}}(\sqrt{P/N}\mathbf{H})$ denotes the K-dimensional capacity region of the interference channel, $C_{\Sigma}^{\mathrm{IC}}(\sqrt{P/N}\mathbf{H})$ denotes the sum-capacity, and R_i denotes the rate corresponding to the i-th transmitter-receiver pair.

5.3 Outer Bound on Capacity Region

First, we present some background on existing outer bounds in literature for interference channels:

5.3.1 Background

The first outer bound on the K-user IFC is obtained by the set of rates $(R_1, \ldots, R_K)$ such that

$$R_i \leq I(X_i; Y_i | X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_K, Q), \quad \forall i, \tag{5.2}$$

for all $p(q)p(x_1|q)p(x_2|q)\cdots p(x_K|q)$, where Q is a time-share variable. This outer bound can be intuitively understood as follows: If each of the receivers can determine the "interference" they observe, they can use it as receiver side-information, which leads to the rates specified by (5.2). In the Gaussian case (with $N = 1$) , this reduces to:

$$R_i \leq \frac{1}{2}\log(1 + h_{ii}^2 P), \quad \forall i. \tag{5.3}$$

An improvement on this outer bound can be obtained using a MAC type approach as studied in [55]. Assuming there exists an $h_{ij} \geq h_{ii}$ for some $j \neq i$, then we have

$$R_i + R_j \leq \frac{1}{2}\log\left(1 + (h_{ij}^2 + h_{jj}^2)P\right). \tag{5.4}$$

This bound is used to demonstrate that the DoF of a Gaussian K-user IFC is upper bounded by $K/2$.

To further improve on this bound, we take the genie aided approach as introduced in [76]. The core idea is to provide additional side information $\widetilde{Y}_1$ to Receiver 1. Assuming that Receiver 1 can determine X_1 with an arbitrarily small probability of error based on Y_1, we desire to provide it with $\widetilde{Y}_1$ so that

a function of X_1, Y_1 and $\widetilde{Y}_1$ results in $\widehat{Y}_2$ that is statistically identical to Y_2. Using this approach, for a two-user Gaussian IFC, an outer bound can be derived as [76]:

$$R_1 + R_2 \leq \frac{1}{2}\log\left((h_{11}^2 P + h_{21}^2 P + 1)\left(\frac{h_{22}^2 P + 1}{\min(h_{21}^2, 1)h_{22}^2 P + 1}\right)\right). \tag{5.5}$$

Generalizing this approach to a Gaussian K-user IFC leads to an explosive growth in the number of parameters to be optimized. We build on this genie-approach to build a "tractable" framework for upper bounding the capacity region of the Gaussian K-user IFC.

5.3.2 Our Outer Bound

In our approach, we create a *genie*-MAC to decode a subset of the messages in the original interference channel. The capacity region of this genie-MAC channel then forms an outer bound on the rate region of the original channel. This genie-MAC technique is a two-step process. The first step is to find a characterization for the genie-MAC receivers, and the second step is to optimize this characterization to obtain the tightest bound of this class.

Consider any permutation function $\pi : \{1, 2, \ldots, K\} \mapsto \{1, 2, \ldots, K\}$, and integers k and m such that $1 \leq k \leq K$ and $m \geq 1$. Define tuples $S = (\pi(1), \ldots, \pi(k))$ and $S^c = (\pi(k+1), \ldots, \pi(K))$. We use X_S to denote the vector $[X_{S(1)}\, X_{S(2)} \quad X_{S(|S|)}]^T$. Now, consider the MAC channel that has X_{S^c}

as side information at the m-antenna receiver and observes the signal

$$\overline{Y} = \mathbf{G}X_S + \overline{Z}, \tag{5.6}$$

where $\overline{Z}$ is i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, for some $\mathbf{G} \in \mathbb{R}^{m\times k}$. Let $C^{\text{MAC}}(\sqrt{P}\mathbf{G}, \mathbf{\Sigma})$ denote the capacity region of this MAC channel and $C_{\Sigma}^{\text{MAC}}(\sqrt{P}\mathbf{G}, \mathbf{\Sigma})$ denote the sum-capacity of this MAC channel. Since the side information is independent of both X_S and $\overline{Z}$, it does not change the capacity region.

Next, we provide the conditions under which the capacity region of this MAC channel form an outer bound on R_S of the original Gaussian IFC.

Lemma 5.1. *Consider any* $\mathbf{T} = [\mathbf{t}_1\, \mathbf{t}_2 \quad \mathbf{t}_k] \in \mathbb{R}^{m\times k}$. *Let* $\mathbf{T}$, $\mathbf{G}$ *and* $\mathbf{\Sigma}$ *be matrices that satisfy the following conditions:*

$$\left(\mathbf{T}^{\dagger}\mathbf{G}\right)^{upper} = \left(\mathbf{H}_S\right)^{upper}, \tag{5.7}$$

$$\mathbf{t}_i^{\dagger}\mathbf{\Sigma}\mathbf{t}_i \leq N, \qquad \forall 1 \leq i \leq k, \tag{5.8}$$

$$\mathbf{\Sigma} \succ 0,$$

where $\mathbf{H}_S$ *is* $|S| \times |S|$ *matrix with entry corresponding to the* i*-th row and* j*-th column as* $h_{S(i),S(j)}$, *i.e.,*

$$\mathbf{H}_S = \begin{bmatrix} h_{S(1),S(1)} & h_{S(1),S(2)} & \dots & h_{S(1),S(k)} \\ h_{S(2),S(1)} & h_{S(2),S(2)} & \dots & h_{S(2),S(k)} \\ \vdots & \vdots & & \vdots \\ h_{S(k),S(1)} & h_{S(k),S(2)} & \dots & h_{S(k),S(k)} \end{bmatrix}.$$

Then,

$$R_S \in C^{MAC}\left(\sqrt{P}\mathbf{G}, \mathbf{\Sigma}\right),$$

i.e., the capacity region of any MAC channel described by (5.6) satisfying the above conditions is an outer bound on the rates R_S for the Gaussian IFC described by (5.1).

Proof. It is sufficient to show the following: If there exists an achievable strategy for the interference channel described by (5.1) to achieve a set of rates $(R_1, R_2, \cdots, R_K)$, i.e., if $(R_1, R_2, \cdots, R_K) \in C^{\mathrm{IC}}(\sqrt{P/N}\mathbf{H})$, then there exists an achievable strategy for the MAC channel described by (5.6) to achieve rates R_S, i.e., $R_S \in C^{\mathrm{MAC}}(\sqrt{P}\mathbf{G}, \mathbf{\Sigma})$. We show that a stricter condition is satisfied. In particular, we prove that the MAC channel can obtain statistically identical (or better) signal as (than) Y_i for all $i \in S$.

Let $\mathbf{D} = \mathbf{T}^\dagger\mathbf{G}$. At the MAC receiver, the signal corresponding to $Y_{S(l)}$ $(1 \le l \le k)$ is obtained sequentially. Consider any step l. Since the messages from transmitters $S(1), S(2), \ldots, S(l-1)$ have been decoded, the receiver can generate signals $X_{S(1)}, X_{S(2)}, \ldots, X_{S(l-1)}$. In addition, the MAC receiver has signals X_{S^c} as side information. Therefore, the MAC receiver can obtain the signal

$$\begin{aligned}
\tilde{Y}_l &= \mathbf{t}_l^\dagger\overline{Y} - \sum_{i=1}^{l-1} d_{l,i}X_{S(i)} + \sum_{i=1}^{l-1} h_{S(l),S(i)}X_{S(i)} + \sum_{i\in S^c} h_{S(l),i}X_i \\
&= \sum_{i=1}^{k} d_{l,i}X_{S(i)} + \mathbf{t}_l^\dagger\overline{Z} - \sum_{i=1}^{l-1} d_{l,i}X_{S(i)} + \sum_{i=1}^{l-1} h_{S(l),S(i)}X_{S(i)} + \sum_{i\in S^c} h_{S(l),i}X_i \\
&= \sum_{i=l}^{k} d_{l,i}X_{S(i)} + \mathbf{t}_l^\dagger\overline{Z} + \sum_{i=1}^{l-1} h_{S(l),S(i)}X_{S(i)} + \sum_{i\in S^c} h_{S(l),i}X_i \\
&= \sum_{i\in S} h_{S(l),i}X_i + \sum_{i\in S^c} h_{S(l),i}X_i + \mathbf{t}_l^\dagger\overline{Z}.
\end{aligned}$$

The last step follows from (5.7), i.e., $d_{l,i} = h_{S(l),S(i)}, \forall i \geq l$. Now, from (5.7), it is clear that $\tilde{Y}_l$ is statistically better than Y_l in the original IFC. Hence, the MAC receiver can decode the message from transmitter $S(l)$ if the receiver $S(l)$ in the original IFC can decode the message from transmitter $S(l)$. This completes the proof of lemma. □

The sum-capacity of the MAC channel is given by

$$C_{\Sigma}^{\text{MAC}}\left(\sqrt{P}\mathbf{G}, \mathbf{\Sigma}\right) = \frac{1}{2}\log\left(|\mathbf{I} + P\mathbf{\Sigma}^{-1}\mathbf{G}\mathbf{G}^{\dagger}|\right).$$

Therefore, we can formulate a minimization problem as follows:

$$\begin{aligned} f^*(\mathbf{H}_S, m) \;&= \inf_{\mathbf{G},\mathbf{\Sigma},\mathbf{T}} \frac{1}{2}\log\left(|\mathbf{I} + P\mathbf{\Sigma}^{-1}\mathbf{G}\mathbf{G}^{\dagger}|\right) \\ \text{such that} \quad & \left(\mathbf{T}^{\dagger}\mathbf{G}\right)^{\text{upper}} = \left(\mathbf{H}_S\right)^{\text{upper}}, \\ & \mathbf{t}_i^{\dagger}\mathbf{\Sigma}\mathbf{t}_i \leq N, \quad \forall 1 \leq i \leq k, \\ & \mathbf{\Sigma} \succ 0. \end{aligned} \tag{5.9}$$

For $m = |S|$, it is clear that the feasible set is non-empty as $\mathbf{G} = \mathbf{H}_S$, $\mathbf{\Sigma} = N\mathbf{I}$ and $\mathbf{T} = \mathbf{I}$ satisfies all the constraints of (5.9). We denote this special case with $m = |S|$ by $f^*(\mathbf{H}_S)$. Now onwards, we assume that $m = |S|$.

From the above analysis, we obtain the following theorem that provides an outer bound on the capacity region of any Gaussian K-user IFC.

Theorem 5.2. *Consider the Gaussian K-user IFC described by (5.1). Then,*

$$C^{IC}\left(\sqrt{P/N}\mathbf{H}\right) \subseteq \left\{(R_1, \ldots, R_K) : \sum_{i \in S} R_i \leq f^*(\mathbf{H}_S), \forall S\right\}, \tag{5.10}$$

i.e., the right hand side (RHS) of (5.10) is an outer bound on the capacity region of this channel.

The above theorem requires the evaluation of the optimization problem given by (5.9). Therefore, next, we derive some results that simplify this optimization problem. In particular, we show that any one of the three parameters can be fixed to identity without affecting the optimal value. The next two lemmas formally state and prove these results.

Lemma 5.3. *Consider the following optimization that results by choosing* $\mathbf{\Sigma} = \mathbf{I}$*:*

$$\begin{aligned} \min_{\mathbf{G},\mathbf{T}} \quad & \frac{1}{2}\log\left(\left|\mathbf{I} + P\mathbf{G}\mathbf{G}^\dagger\right|\right) \\ \textit{such that} \quad & \left(\mathbf{T}^\dagger\mathbf{G}\right)^{upper} = \left(\mathbf{H}_S\right)^{upper}, \\ & \mathbf{t}_i^\dagger\mathbf{t}_i \leq N, \quad \forall 1 \leq i \leq k. \end{aligned} \tag{5.11}$$

Then, the optimal value of this problem is $f^*(\mathbf{H}_S)$.

Proof. Consider any feasible set of parameters $\mathbf{G}$, $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^\dagger$ and $\mathbf{T}$ for the optimization problem given by (5.9). Let $\hat{\mathbf{G}} = \mathbf{A}^{-1}\mathbf{G}$ and $\hat{\mathbf{T}} = \mathbf{A}^\dagger\mathbf{T}$. Now, we have the following:

$$\hat{\mathbf{T}}^\dagger\hat{\mathbf{G}} = \mathbf{T}^\dagger\mathbf{A}\mathbf{A}^{-1}\mathbf{G} = \mathbf{T}^\dagger\mathbf{G},$$

$$\hat{\mathbf{T}}^\dagger\hat{\mathbf{T}} = \mathbf{T}^\dagger\mathbf{A}\mathbf{A}^\dagger\mathbf{T} = \mathbf{T}^\dagger\mathbf{\Sigma}\mathbf{T}.$$

Therefore, $\hat{\mathbf{G}}$ and $\hat{\mathbf{T}}$ form a feasible set for the optimization problem given by (5.11). Furthermore, the objective value remains the same due to the following:

$$\begin{aligned} |\mathbf{I} + P\hat{\mathbf{G}}\hat{\mathbf{G}}^\dagger| &= |\mathbf{I} + P\mathbf{A}^{-1}\mathbf{G}\mathbf{G}^\dagger(\mathbf{A}^\dagger)^{-1}| \\ &= |\mathbf{I} + P(\mathbf{A}^\dagger)^{-1}\mathbf{A}^{-1}\mathbf{G}\mathbf{G}^\dagger| \\ &= |\mathbf{I} + P\mathbf{\Sigma}^{-1}\mathbf{G}\mathbf{G}^\dagger|. \end{aligned}$$

This completes the proof. □

Lemma 5.4. *Consider the optimization problem given by (5.9). Now, consider the two sub-problems that result from choosing either* $\mathbf{T} = \mathbf{I}$ *or* $\mathbf{G} = \mathbf{I}$*. Then, each of these sub-problems has optimal value* $f^*(\mathbf{H}_S)$*.*

Proof. Case-I ($\mathbf{T} = \mathbf{I}$): Consider any feasible set of parameters $\hat{\mathbf{G}}$ and $\hat{\mathbf{T}}$ for the optimization problem given by (5.11). Let ϵ be an arbitrary real number such that $0 < \epsilon < 1$. Let $\mathbf{G} = \hat{\mathbf{T}}^\dagger\hat{\mathbf{G}}$ and $\mathbf{\Sigma} = \epsilon N\mathbf{I} + (1-\epsilon)\hat{\mathbf{T}}^\dagger\hat{\mathbf{T}}$. It is fairly straightforward to check that these parameters are feasible for the sub-problem. Furthermore, as the objective function is continuous, the objective value approaches that of the original problem as $\epsilon \to 0$.

Case-II ($\mathbf{G} = \mathbf{I}$): Consider any feasible set of parameters $\hat{\mathbf{G}}$ and $\hat{\mathbf{T}}$ for the optimization problem given by (5.11). Let ϵ be an arbitrary real number such that $0 < \epsilon < 1$. Let $\mathbf{T} = \hat{\mathbf{G}}^\dagger\hat{\mathbf{T}}$ and $\mathbf{\Sigma} = (\epsilon\mathbf{I} + \hat{\mathbf{G}}^\dagger\hat{\mathbf{G}})^{-1}$. Again, it is fairly straightforward to check that these parameters are feasible for the sub-problem, and the objective value approaches that of the original problem as $\epsilon \to 0$. □

Next, we compare this outer bound expression with other techniques in literature. It is fairly simple to see that this bound incorporates receiver cooperation as a special case. In particular, by choosing the genie-MAC channel matrix $\mathbf{G}$ to be the same as the channel gains in the original IFC, the receiver cooperative bound can be obtained. A multiple-access type outer bound as studied in [19, 103] is also a special case of this bound. A conventional MAC-type bound corresponds to the case when S is a set of the form $\{i, j\}$. It is perhaps not as straightforward to see that this is, in fact, a genie-aided outer bound. If we were to choose a subset of the rows of the matrix $\mathbf{G}$ to match those in the original interference channel definition, then the remaining rows of $\mathbf{G}$ along with X_{S^c} represent a "vector genie" provided to enable all messages to be decoded in the system. This bound does not capture all genie-aided bounds in the two-user setting.

Although it captures many existing bounding techniques for the interference channel, the optimization problem in (5.11) does not necessarily lend itself to a straightforward solution. Furthermore, to evaluate the bound on the sum of a set of rates, we need to consider all possible orderings of tuples S resulting from this set. In the next section, we show that analyzing this bound is indeed tractable in the case of degraded Gaussian K-user IFCs and yields closed-form expressions for sum-capacity. However, this analysis does not follow from standard convex optimization techniques. Instead, we provide a novel construction-based proof.

5.4 Sum Capacity of Degraded Gaussian IFCs

In this section, we determine the sum-capacity of a class of Gaussian K-user interference channels called the *degraded* IFCs. This class of channels is interesting for multiple reasons: (i) It does not belong to known classes including "weak" and "strong" interference channels. (ii) It corresponds to a case where users see highly dependent signals, which may very well be the case in certain wireless or wired network settings.

For a stochastically degraded discrete memoryless IFC, there exist probability mass functions $q_i(y_i|y_{i-1})$ for $2 \leq j \leq n$ such that:

$$p(y_j|x_1^n) = p(y_1|x_1^n) \prod_{i=2}^{n} q_i(y_i|y_{i-1}),$$

for all $2 \leq j \leq n$. This definition extends naturally to the class of degraded Gaussian IFCs as well. In essence, there is an ordering among receivers, with Receiver 1 observing the "least corrupted" signal and Receiver K observing the "most degraded". It is easy to see that this implies that all degraded Gaussian IFC channel matrices can be expressed as $\mathbf{H} = \mathbf{a}\mathbf{b}^\dagger$, where $\mathbf{a} = [a_1\, a_2\, \ldots\, a_K]^T$ and $\mathbf{b} = [b_1\, b_2\, \ldots\, b_K]^T$. Without loss of generality, we assume $a_1^2 \leq a_2^2 \leq \ldots \leq a_K^2$, and $P = N = 1$. Note that in this ordering Receiver 1 is "most degraded".

5.4.1 Achievability

We utilize a successive interference cancellation scheme for achievability. Each transmitter uses Gaussian codewords to encode its message. The

i-th receiver decodes the messages from transmitters $1, 2, \ldots, i$ in this order. Since i-th receiver has a (statistically) better received signal than receivers $1, 2, \ldots, i-1$, the message at i-th transmitter can be encoded at rate

$$R_i = \frac{1}{2}\log\left(1 + \frac{a_i^2 b_i^2}{a_i^2\left(\sum_{j=i+1}^{K} b_j^2\right) + 1}\right) \tag{5.12}$$

such that all receivers $i, i+1, \ldots, K$ can decode it with decaying probability of error. Since this is a well-known technique, we do not provide further details. From (5.12), the achievable sum-rate using this SIC scheme can be expressed as

$$\begin{aligned}
\sum_{i=1}^{K} R_i &= \frac{1}{2}\sum_{i=1}^{K}\log\left(\frac{a_i^2\left(\sum_{j=i}^{K} b_j^2\right) + 1}{a_i^2\left(\sum_{j=i+1}^{K} b_j^2\right) + 1}\right), \\
&= \frac{1}{2}\log\left(\frac{\prod_{i=1}^{K}\left(a_i^2\left(\sum_{j=i}^{K} b_j^2\right) + 1\right)}{\prod_{i=1}^{K}\left(a_{i-1}^2\left(\sum_{j=i}^{K} b_j^2\right) + 1\right)}\right), \\
&= \frac{1}{2}\sum_{i=1}^{K}\log\left(1 + \frac{(a_i^2 - a_{i-1}^2)\left(\sum_{j=i}^{K} b_j^2\right)}{a_{i-1}^2\left(\sum_{j=i}^{K} b_j^2\right) + 1}\right),
\end{aligned} \tag{5.13}$$

where $a_0 = 0$ is introduced for notational convenience.

5.4.2 Outer Bound

The non-trivial step is to obtain a matching outer bound on sum-rate. We apply the general technique developed in Section 5.3 to obtain the outer bound. As discussed before, it is very hard to evaluate these bounds (especially in closed-form) for general IFCs, but the degraded structure of the channel can be exploited as shown next.

Consider the optimization problem given by (5.11) for the tuple $S = (1, 2, \ldots, K)$. Solving this is equivalent to showing the existence of feasible $\mathbf{G}$ and $\mathbf{T}$ that evaluates to the RHS of (5.13). Now, consider the following construction for $\mathbf{G}$ and $\mathbf{T}$. Given any i such that $1 \leq i \leq K$, let

$$c_i = \sqrt{a_i^2 - a_{i-1}^2}, \tag{5.14}$$

and $\mathbf{c} = [c_1\ c_2\ \ldots\ c_K]^T$. We use the following iterative construction to obtain a upper-triangular matrix $\mathbf{T}$ (lower-triangular $\mathbf{T}^\dagger$):

$$\mathbf{t}_i = \frac{a_{i-1}}{a_i}\mathbf{t}_{i-1} + \frac{c_i}{a_i}\mathbf{e}_i, \quad \forall i, \tag{5.15}$$

where $\mathbf{t}_0 = \mathbf{0}$ and $\mathbf{e}_i$ is the unit-vector along i-th dimension. The entry corresponding to the i-th row and j-th column of $\mathbf{G}$ is chosen as

$$g_{i,j} = c_i b_j d_{i,j}, \quad \forall i, j, \tag{5.16}$$

where $d_{i,j}$ parameters are introduced here for the first time. We fix $d_{i,j} = 1$ for any $i \leq j$. The choice of remaining parameters ($d_{i,j}$ for $i > j$) are discussed later. Irrespective of these remaining parameters, the above construction has the following property.

Lemma 5.5. *Consider any* $\mathbf{G}$ *and* $\mathbf{T}$ *given above. Then, it belongs to the feasible set corresponding to the optimization problem given by (5.11), i.e.,*

$$\left(\mathbf{T}^\dagger\mathbf{G}\right)^{upper} = (\mathbf{H})^{upper},$$
$$\mathbf{t}_i^\dagger\mathbf{t}_i \leq 1, \qquad \forall 1 \leq i \leq K.$$

Proof. First, for all i, we show that $\mathbf{t}_i^\dagger \mathbf{t}_i = 1$ by induction. Since $\mathbf{t}_1 = \mathbf{e}_1$, we have $\mathbf{t}_1^\dagger \mathbf{t}_1 = 1$. By construction, we have $\mathbf{t}_{i-1}^\dagger \mathbf{e}_i = 0$. Suppose that $\mathbf{t}_{i-1}^\dagger \mathbf{t}_{i-1} = 1$ for some i. Then, from (5.15) and (5.14), we have

$$\begin{aligned} \mathbf{t}_i^\dagger \mathbf{t}_i &= \frac{a_{i-1}^2}{a_i^2} \mathbf{t}_{i-1}^\dagger \mathbf{t}_{i-1} + \frac{c_i^2}{a_i^2}, \\ &= \frac{a_{i-1}^2}{a_i^2} + \frac{a_i^2 - a_{i-1}^2}{a_i^2}, \\ &= 1. \end{aligned} \tag{5.17}$$

Next, for all i, we show that $\mathbf{t}_i^\dagger \mathbf{c} = a_i$ by induction. Since $\mathbf{t}_1 = \mathbf{e}_1$, we have $\mathbf{t}_1^\dagger \mathbf{c} = a_1$. Suppose that $\mathbf{t}_{i-1}^\dagger \mathbf{c} = a_{i-1}$ for some i. Then, from (5.15) and (5.14), we have

$$\begin{aligned} \mathbf{t}_i^\dagger \mathbf{c} &= \frac{a_{i-1}}{a_i} \mathbf{t}_{i-1}^\dagger \mathbf{c} + \frac{c_i}{a_i} c_i, \\ &= \frac{a_{i-1}^2}{a_i} + \frac{a_i^2 - a_{i-1}^2}{a_i}, \\ &= a_i. \end{aligned} \tag{5.18}$$

Last, for all $i \leq j$, using lower-triangular property of $\mathbf{T}^\dagger$ and (5.18), we show that the (i, j)-th entry of $\mathbf{T}^\dagger \mathbf{G}$ is equal to $h_{i,j}$:

$$\begin{aligned} (\mathbf{T}^\dagger \mathbf{G})_{i,j} &= \mathbf{t}_i^\dagger b_j [d_{1,j} c_1 \; d_{2,j} c_2 \; \ldots \; d_{K,j} c_K]^T, \\ &= \mathbf{t}_i^\dagger [c_1 \; c_2 \; \ldots \; c_K]^T b_j, \quad \forall i \leq j, \\ &= a_i b_j, \quad \forall i \leq j. \end{aligned} \tag{5.19}$$

With (5.17) and (5.19), the proof is complete. □

Next, we show that parameters $d_{i,j}$ (for $i > j$) exist such that (5.11) evaluates to RHS of (5.13). For this, we consider a lower-triangular matrix $\mathbf{V}$ with unit diagonal entries. Let (i,j)-th entry of $\mathbf{V}$ be denoted by $v_{i,j}$. Define $\mathbf{F} = \mathbf{I} + \mathbf{G}\mathbf{G}^\dagger$. Therefore, from (5.16), the (i,j)-th entry of $\mathbf{VF}$ is

$$\begin{aligned}(\mathbf{VF})_{i,j} &= \sum_{m=1}^{i}\left(v_{i,m}\left(\delta_{m,j} + \sum_{n=1}^{K} g_{m,n}g_{j,n}\right)\right), \\ &= \sum_{m=1}^{i} v_{i,m}\delta_{m,j} + c_j \sum_{n=1}^{K}\left(b_n^2 d_{j,n}\left(\sum_{m=1}^{i} v_{i,m}c_m d_{m,n}\right)\right). \quad (5.20)\end{aligned}$$

Now, suppose that, for all $i \geq 2$ and $n \leq i-1$, the parameters are such that

$$\sum_{m=1}^{i} v_{i,m}c_m d_{m,n} = 0, \quad \forall i \geq 2, n \leq i-1. \quad (5.21)$$

Then, for all i and $j \leq i$, substituting (5.21) and $d_{i,j} = 1$ for any $i \leq j$ in (5.20) , we obtain

$$(\mathbf{VF})_{i,j} = v_{i,j} + c_j \sum_{n=i}^{K}\left(b_n^2\left(\sum_{m=1}^{i} v_{i,m}c_m\right)\right), \forall i, j \leq i. \quad (5.22)$$

For the set of values given by

$$v_{i,j} = \frac{-c_i c_j \sum_{n=i}^{K} b_n^2}{\left(\sum_{m=1}^{i-1} c_m^2\right)\left(\sum_{n=i}^{K} b_n^2\right) + 1}, \forall j < i, \quad (5.23)$$

from (5.22), we have $(\mathbf{VF})_{i,j} = 0$ for all $j < i$ (i.e., $\mathbf{VF}$ is upper-triangular)

and

$$\begin{aligned}(\mathbf{VF})_{i,i} &= 1 + c_i \sum_{n=i}^{K} \left(b_n^2 \left(\sum_{m=1}^{i-1} v_{i,m} c_m + c_i \right) \right), \\ &= 1 + \frac{c_i^2 \left(\sum_{n=i}^{K} b_n^2 \right)}{\left(\sum_{m=1}^{i-1} c_m^2 \right) \left(\sum_{n=i}^{K} b_n^2 \right) + 1}, \\ &= 1 + \frac{(a_i^2 - a_{i-1}^2) \left(\sum_{j=i}^{K} b_j^2 \right)}{a_{i-1}^2 \left(\sum_{j=i}^{K} b_j^2 \right) + 1}, \quad \forall i. \end{aligned} \tag{5.24}$$

Substituting (5.23) in (5.21), we obtain

$$c_i \left(\frac{- \left(\sum_{m=1}^{n} c_m^2 + \sum_{m=n+1}^{i-1} c_m^2 d_{m,n} \right) \left(\sum_{j=i}^{K} b_j^2 \right)}{\left(\sum_{m=1}^{i-1} c_m^2 \right) \left(\sum_{j=i}^{K} b_j^2 \right) + 1} + d_{i,n} \right) = 0, \tag{5.25}$$

for all $i \geq 2$ and $n \leq i-1$. For any given n, it is clear that we can choose $d_{i,n}$ for all $i > n$, such that (5.25) is satisfied for all $i > n$. This directly follows form the fact these are linear equations in $d_{i,n}$ with same number of variables as equations. Therefore, we have a construction that satisfies the assumption in (5.21). For understanding the structure of the genie-MAC channel matrix $\mathbf{G}$ given by (5.16), it is useful to evaluate $d_{i,n}$ that satisfies (5.25). From (5.16), we can express $d_{i,n}$ for $i > n$ as follows: for all $1 \leq n \leq K$

$$\begin{aligned} d_{n+1,n} &= \frac{\left(\sum_{m=1}^{n} c_m^2 \right) \left(\sum_{j=i}^{K} b_j^2 \right)}{\left(\sum_{m=1}^{n} c_m^2 \right) \left(\sum_{j=i}^{K} b_j^2 \right) + 1}, \\ d_{i,n} &= \frac{\left(\sum_{m=1}^{n} c_m^2 + \sum_{m=n+1}^{i-1} c_m^2 d_{m,n} \right) \left(\sum_{j=i}^{K} b_j^2 \right)}{\left(\sum_{m=1}^{i-1} c_m^2 \right) \left(\sum_{j=i}^{K} b_j^2 \right) + 1}, \qquad \forall i > n+1. \end{aligned}$$

Now, for the above construction, $\mathbf{VF}$ is upper-triangular and $|\mathbf{V}| = 1$.

Therefore, from (5.24), we have

$$\begin{aligned}\frac{1}{2}\log|\mathbf{I}+\mathbf{G}\mathbf{G}^{\dagger}| &= \frac{1}{2}\log|\mathbf{F}| \\ &= \frac{1}{2}\log|\mathbf{V}\mathbf{F}| \\ &= \frac{1}{2}\log\prod_{i=1}^{K}(\mathbf{V}\mathbf{F})_{i,i}, \\ &= \frac{1}{2}\sum_{i=1}^{K}\log\left(1+\frac{(a_i^2-a_{i-1}^2)\left(\sum_{j=i}^{K}b_j^2\right)}{a_{i-1}^2\left(\sum_{j=i}^{K}b_j^2\right)+1}\right),\end{aligned}$$

which exactly matches the achievable sum-rate in (5.13).

5.4.3 Sum Capacity

The above analysis establishes the sum-capacity of the class of degraded Gaussian K-user IFCs. We summarize this result in the following theorem.

Theorem 5.6. *Consider any degraded Gaussian K-user interference channel with $\mathbf{H} = \mathbf{a}\mathbf{b}^{\dagger}$, where $\mathbf{a} = [a_1\ a_2\ \ldots\ a_K]^T$ and $\mathbf{b} = [b_1\ b_2\ \ldots\ b_K]^T$. Let $a_1^2 \leq a_2^2 \leq \ldots \leq a_K^2$ and $a_0 = 0$. Then, the sum-capacity of this channel is given by*

$$C_{\Sigma}^{IC}\left(\sqrt{P/N}\mathbf{H}\right) = \frac{1}{2}\sum_{i=1}^{K}\log\left(1+\frac{(a_i^2-a_{i-1}^2)\left(\sum_{j=i}^{K}b_j^2\right)P}{a_{i-1}^2\left(\sum_{j=i}^{K}b_j^2\right)P+N}\right).$$

Remark 5.1. *This class of channels have degree of freedom equal to 1. The degree of freedom can be obtained in a straightforward manner as the K-th receiver can decode messages from all transmitters. However, this approach does not give the required tight outer bound on sum-rate.*

5.5 MIMO Unit-Rank Interference Channels

In this section, we extend the sum-capacity results to MIMO unit-rank interference channels, which are sometimes referred to as *keyhole* channels as mentioned in the introduction. Consider a MIMO unit-rank Gaussian IFC described by

$$Y_{i,n}[t] = a_{i,n} \sum_{j=1}^{K} \left(\sum_{m=1}^{M_j} b_{j,m} X_{j,m}[t] \right) + Z_{i,n}[t], \tag{5.26}$$

for all $1 \leq n \leq N_i, 1 \leq i \leq K$. Here, N_i is the number of antennas at i-th receiver and M_j is the number of antennas at j-th transmitter. The power constraint at the j-th transmitter is $\mathbb{E}[\|X_j\|^2] \leq P$, and the zero-mean i.i.d. Gaussian noise at all receivers have covariance $N\mathbf{I}$.

The following theorem establishes an equivalence to the SISO scenario.

Theorem 5.7. *The capacity region of the MIMO unit-rank Gaussian IFC in (5.26) is identical to the capacity region of the SISO degraded Gaussian IFC described by*

$$\hat{Y}_i[t] = \|\mathbf{a}_i\| \sum_{j=1}^{K} \|\mathbf{b}_j\| \hat{X}_j[t] + \hat{Z}_i[t], \quad \forall 1 \leq i \leq K. \tag{5.27}$$

Proof. First, we prove the relatively straightforward result that the capacity region of the SISO Gaussian IFC is a subset of the capacity region of the MIMO Gaussian IFC. Consider the following transformation at the transmitters of the MIMO Gaussian IFC:

$$\tilde{X}_{j,m}[t] = \frac{b_{j,m}}{\|\mathbf{b}_j\|} \hat{X}_j[t]. \tag{5.28}$$

This transformation in (5.28) satisfies the power constraints for the MIMO Gaussian IFC due to the following:

$$\begin{aligned}\sum_{m=1}^{M_j} \tilde{X}_{j,m}^2[t] &= \sum_{m=1}^{M_j} \frac{b_{j,m}^2}{\|\mathbf{b}_j\|^2} \hat{X}_j^2[t], \\ &= \hat{X}_j^2[t]. \end{aligned} \tag{5.29}$$

Next, consider the following transformation at the receivers of the MIMO Gaussian IFC:

$$\begin{aligned}\tilde{Y}_i[t] &= \sum_{n=1}^{N_i} \frac{a_{i,n}}{\|\mathbf{a}_j\|} Y_{i,n}[t], \\ &= \|\mathbf{a}_i\| \sum_{j=1}^{K} \|\mathbf{b}_j\| \hat{X}_j[t] + \sum_{n=1}^{N_i} \frac{a_{i,n}}{\|\mathbf{a}_j\|} Z_{i,n}[t]. \end{aligned} \tag{5.30}$$

Since the additive noise in (5.30) is $\mathcal{N}(0, N)$ and transmitter power constraint is clearly satisfied due to (5.29), the above transformations show that the capacity region of the SISO Gaussian IFC is a subset of the capacity region of the MIMO Gaussian IFC.

Next, we prove that the capacity region of the MIMO Gaussian IFC is a subset of the capacity region of the SISO Gaussian IFC. Consider the following transformation at the transmitters of the SISO Gaussian IFC:

$$\tilde{X}_j[t] = \sum_{m=1}^{M_j} \frac{b_{j,m}}{\|\mathbf{b}_j\|} X_{j,m}[t].$$

Using CauchySchwarz inequality, we obtain

$$\begin{aligned}\tilde{X}_j^2[t] &= \frac{\left(\sum_{m=1}^{M_j} b_{j,m} X_{j,m}[t]\right)^2}{\|\mathbf{b}_j\|^2}, \\ &\leq \frac{\left(\sum_{m=1}^{M_j} b_{j,m}^2\right)\left(\sum_{m=1}^{M_j} X_{j,m}^2[t]\right)}{\|\mathbf{b}_j\|^2}. \\ &= \sum_{m=1}^{M_j} X_{j,m}^2[t].\end{aligned}$$

Therefore, this transformation satisfies the power constraints for the SISO Gaussian IFC. Next, consider the following transformation at the receivers of the SISO Gaussian IFC:

$$\tilde{Y}_{i,n}[t] = \frac{a_{i,n}}{\|\mathbf{a}_j\|}\hat{Y}_i[t] + W_{i,n}, \tag{5.31}$$

where $(W_{i,1}, W_{i,2}, \ldots, W_{i,N_i})$ is $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_i)$ (and independent of all other random variables). These received signals are statistically identical to the received signals of MIMO Gaussian IFC if

$$\mathbf{\Sigma}_i = \mathbf{I} - \frac{\mathbf{a}_i \mathbf{a}_i^\dagger}{\|\mathbf{a}_i\|^2} \tag{5.32}$$

and these are positive semi-definite matrices. The following shows that these are positive semi-definte matrices:

$$\begin{aligned}\mathbf{x}^\dagger \mathbf{\Sigma}_i \mathbf{x} &= \mathbf{x}^\dagger \mathbf{x} - \frac{\mathbf{x}^\dagger \mathbf{a}_i \mathbf{a}_i^\dagger \mathbf{x}}{\|\mathbf{a}_i\|^2} \\ &= \mathbf{x}^\dagger \mathbf{x} - \frac{\mathrm{tr}\left(\mathbf{x}\mathbf{x}^\dagger \mathbf{a}_i \mathbf{a}_i^\dagger\right)}{\|\mathbf{a}_i\|^2} \\ &\geq \mathbf{x}^\dagger \mathbf{x} - \frac{\mathrm{tr}\left(\mathbf{x}\mathbf{x}^\dagger\right)\mathrm{tr}\left(\mathbf{a}_i \mathbf{a}_i^\dagger\right)}{\|\mathbf{a}_i\|^2} \\ &= 0.\end{aligned} \tag{5.33}$$

Thus, the above transformations show that the capacity region of the MIMO Gaussian IFC is a subset of the capacity region of the SISO Gaussian IFC. This completes the proof. □

This immediately leads to the following corollary.

Corollary 5.8. *The sum-capacity of the MIMO unit-rank interference channel described in (5.26) is given by*

$$\frac{1}{2}\sum_{i=1}^{K}\log\left(1+\frac{\left(\|\mathbf{a}_i\|^2-\|\mathbf{a}_{i-1}\|^2\right)\left(\sum_{j=i}^{K}\|\mathbf{b}_j\|^2\right)P}{\|\mathbf{a}_{i-1}\|^2\left(\sum_{j=i}^{K}\|\mathbf{b}_j\|^2\right)P+N}\right).$$

Proof. The proof follows from Theorem 5.6 and Theorem 5.7. □

Chapter 6

Robust Weighted-Sum Rate Maximization in Multiple Antenna Interference Networks

6.1 Introduction

Limited progress[1] has been made in obtaining the capacity region of K-user Gaussian interference channels (GIFC) whereas significant advances have been made in understanding *degrees-of-freedom*[2] of these channels. In [18,47], the degrees-of-freedom (DoF) optimality of interference alignment for time-varying or frequency-selective SISO and MIMO channels. With limited number of dimensions (time or frequency) and/or finite power levels, in general, interference alignment is not capacity-region optimal. Even for two-user GIFCs, it is well known that more involved schemes like Han-Kobayashi scheme are required to be within one-bit of the capacity region [41]. However, these DoF results strongly motivate achievable schemes with linear transmit precoding, linear receive filtering and point-to-point coding. Additionally, such achievable schemes are greatly motivated from an implementation perspective in cellular standards such as LTE-A - CoMP: Coordinated Multi-Point TX/RX [1] and

[1] We refer readers to Chapter 5 for details.

[2] Degrees-of-freedom is the pre-log factor in sum rate asymptotically as per-user power is increased to infinity.

802.16m - Multi-BS MIMO [6]. Further, these standards motivate the need to develop distributed schemes based on limited (imperfect) channel state information (CSI) at the transmitters.

While focusing on linear schemes, the resulting *robust* transmit precoder designs can be formulated as optimization problems in a straightforward manner. Solutions to such optimization problems have been widely studied in the context of MIMO broadcast channels [75, 127] and MISO interference networks [118], and have proved to be of immediate applicability in practice. However, there is very limited work on MIMO interference networks under channel uncertainty. Majority of existing literature on MIMO interference networks focuses on sum-rate or MMSE objectives in the presence of perfect CSI [45, 98, 100]. There has been recent work on weighted-sum rate objective in [105] and rank minimization approach in [97], however, all these results are limited to perfect CSI.

While dealing with interference networks, the optimization problems are known to be hard. In particular, the weighted sum rate optimization problem even in the SISO interference channel with perfect CSI was shown to be NP hard in [82]. Thus, globally optimal precoder design for the weighted sum rate metric over MIMO GIFCs cannot be determined efficiently and hence sub-optimal algorithms must be employed. However, instead of resorting to ad-hoc algorithms, a systematic approach that provides partial guarantees is that of formulating sub-problems that can be solved optimally, either in closed-form or through standard convex optimization formulations such as SDP or MAX-

DET [16,122]. This approach leads to iterative algorithms that are guaranteed to converge. Applying this approach is highly non-trivial and requires careful utilization of the structure present in the optimization problems.

6.1.1 Our Contributions

We consider weighted-sum rate and max-min rate objectives, and two decoding schemes: single-stream decoding and single-user decoding. Under channel uncertainty, we consider bounded-error models for quantization and robust counterparts of the objectives. Our main contributions are the following:

- We derive precoder design algorithms using a provably convergent iterative approach for all these scenarios.
- We extend the genie-MAC formulation in Chapter 5 to obtain outer bounds under channel uncertainty. We develop computable bounds using an alternating optimization approach.

6.1.2 Organization

In Section 6.2, we present the system model. In Section 6.3, we provide relevant known results. In Section 6.4, we describe the robust weighted-sum rate problem and our results for single-stream decoding. We extend these results to single-user decoding in Section 6.5 and to robust max-min rate objective in Section 6.6. In Section 6.7, we provide our results on genie-MAC outer bounds. Finally, in Section 6.8, we present numerical results.

6.2 System Model

We consider K-user MIMO Gaussian interference channels consisting of K transmitter-receiver pairs, labeled $1, 2, \ldots, K$. The j-th transmitter has M_j antennas and the i-th receiver has N_i antennas.[3] Each transmitter has independent messages for the corresponding receiver. We assume a discrete-time channel model as described next. At time $t, t \in \mathbb{Z}_+$, the input-output relations that describe the interference channel are:

$$\mathbf{y}_i[t] = \sum_{j=1}^{K} \mathbf{H}_{i,j}\mathbf{x}_j[t] + \mathbf{z}_i[t], \tag{6.1}$$

where $\mathbf{x}_j[t] \in \mathbb{C}^{M_j \times 1}$ is the signal transmitted by the j-th transmitter, $\mathbf{H}_{i,j} \in \mathbb{C}^{N_i \times M_j}$ is the constant channel matrix from j-th transmitter to i-th receiver, $\mathbf{z}_i[t] \in \mathbb{C}^{N_i \times 1}$ is the additive complex Gaussian $\mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$ noise at i-th receiver, and $\mathbf{y}_i[t] \in \mathbb{C}^{N_i \times 1}$ is the signal received at the i-th receiver. For simplicity, we suppress the time index t henceforth. The power constraint at every transmitter is $\mathbb{E}[\|\mathbf{x}_j\|^2] \leq P, \forall j$.

6.3 Preliminaries

Lemma 6.1 (Shur Complement Lemma)**.** *Let* $\mathbf{A}$*,* $\mathbf{B}$ *and* $\mathbf{C}$ *be given matrices, and* $\mathbf{X}$ *be the symmetric matrix given by*

$$\mathbf{X} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\dagger & \mathbf{C} \end{bmatrix}.$$

[3]Note that these antennas can model time/frequency selective channels by appropriately choosing block-diagonal channel matrices.

If $\mathbf{A}$ *is positive definite, then* $\mathbf{X}$ *is positive semidefinite if and only if*

$$\mathbf{C} - \mathbf{B}^\dagger \mathbf{A}^{-1} \mathbf{B} \succeq 0.$$

The following lemma with important consequences in robust optimization has been proved in [37].

Lemma 6.2. *Let* $\mathbf{A}$, $\mathbf{B}$ *and* $\mathbf{C}$ *be given matrices, with* $\mathbf{A} = \mathbf{A}^\dagger$. *Then, the relation*

$$\mathbf{A} \succeq \mathbf{B}^\dagger \mathbf{D} \mathbf{C} + \mathbf{C}^\dagger \mathbf{D}^\dagger \mathbf{B}, \forall \mathbf{D} : \|\mathbf{D}\|_2 \leq \epsilon$$

is valid, if and only if

$$\exists \lambda \geq 0, \begin{bmatrix} \mathbf{A} - \lambda \mathbf{C}^\dagger \mathbf{C} & -\epsilon \mathbf{B}^\dagger \\ -\epsilon \mathbf{B} & \lambda \mathbf{I} \end{bmatrix} \succeq 0.$$

The following is a useful lemma from [28] that allows us to introduce auxiliary variables to obtain optimally solvable sub-problems.

Lemma 6.3. *Consider the function* $f(s) = -es + \log s + 1$ *for any given* $e > 0$. *Then,*

$$\max_{s \in \mathbb{R}_+} f(s) = \log \left(\frac{1}{e} \right),$$

with the optimum value $s^{opt} = 1/e$.

A matrix version of the above lemma is as follows:

Lemma 6.4. *Let* d *be any integer and* $\mathbf{E} \in \mathbb{C}^{d \times d}$ *be any matrix such that* $\mathbf{E} \succ 0$ *and* $|\mathbf{E}| \leq 1$. *Consider the function* $f(\mathbf{S}) = -\operatorname{tr}(\mathbf{S}\mathbf{E}) + \log |\mathbf{S}| + d$.

Then,

$$\max_{\mathbf{S}\in\mathbb{C}^{d\times d},\,\mathbf{S}\succ 0} f(\mathbf{S}) = \log|\mathbf{E}^{-1}|,$$

with the optimum value $\mathbf{S}^{opt} = \mathbf{E}^{-1}$.

This lemma is obtained via Fenchel conjugate arguments [15].

6.4 Robust Weighted-Sum Rate Maximization

We focus on a simple communication scheme that uses linear transmit precoders and linear receive filters so that i-th transmitter-receiver pair can transmit and receive d_i streams, which are encoded and decoded independently using point-to-point schemes. We denote the independently encoded symbols at j-th transmitter by $\hat{\mathbf{x}}_j \in \mathbb{C}^{d_j\times 1}$, the precoder at j-th transmitter by $\mathbf{V}_j \in \mathbb{C}^{M_j\times d_j}$, and the filter at i-th receiver by $\mathbf{G}_i \in \mathbb{C}^{N_i\times d_i}$. Now, using (6.1), the output of the i-th receiver filter can be expressed as

$$\hat{\mathbf{y}}_i = \sum_{j=1}^{K} \mathbf{G}_i^{\dagger}\mathbf{H}_{i,j}\mathbf{V}_j\hat{\mathbf{x}}_j + \mathbf{G}_i^{\dagger}\mathbf{z}_i. \tag{6.2}$$

From (6.2), the received signal corresponding to the l-th stream is given by

$$\hat{y}_{i,l} = \sum_{j=1}^{K}\sum_{k=1}^{d_j} \mathbf{g}_{i,l}^{\dagger}\mathbf{H}_{i,j}\mathbf{v}_{j,k}\hat{x}_{j,k} + \mathbf{g}_{i,l}^{\dagger}\mathbf{z}_i. \tag{6.3}$$

Now, the achievable rate corresponding to this stream can be expressed as

$$R_{i,l} = \max_{\mathbf{g}_{i,l}} \quad \log\left(1 + \frac{|\mathbf{g}_{i,l}^{\dagger}\mathbf{H}_{i,i}\mathbf{v}_{i,l}|^2}{\sum_{(j,k)\neq(i,l)}|\mathbf{g}_{i,l}^{\dagger}\mathbf{H}_{i,j}\mathbf{v}_{j,k}|^2 + \sigma^2\|\mathbf{g}_{i,l}\|^2}\right). \tag{6.4}$$

The first problem of interest is the precoder design that maximizes the weighted sum-rate, which can be formulated as

$$\begin{aligned} \max_{\{\mathbf{V}_j\}} \quad & \sum_{i=1}^{K} \sum_{l=1}^{d_i} w_i R_{i,l} \\ \text{s.t.} \quad & \|\mathbf{V}_j\|_F^2 \leq P, \forall j, \end{aligned} \tag{6.5}$$

where $w_i \in \mathbb{R}_+$ are given weights.

Remark 6.1. *The motivation behind considering weighted-sum rate is that network resource allocation is often performed by adapting these weights over time (at a larger time-scale). From a physical-layer perspective, these weights can be considered as given constants.*

Given the presence of channel uncertainty at the transmitters in practice, a more relevant and difficult problem of interest is a robust counterpart of (6.5) in the presence of bounded channel errors, which is explained next. We assume that every channel matrix can be expressed as

$$\mathbf{H}_{i,j} = \hat{\mathbf{H}}_{i,j} + \mathbf{\Delta}_{i,j}, \tag{6.6}$$

where $\hat{\mathbf{H}}_{i,j}$ is the channel estimate known to the transmitters, and $\mathbf{\Delta}_{i,j}$ is the unknown error with $\|\mathbf{\Delta}_{i,j}\|_F \leq \epsilon_{i,j}$ for some given $\epsilon_{i,j} \geq 0$. A natural worst-case formulation of the optimization problem in (6.5) under channel uncertainty in (6.6) is the following:

$$\max_{\{\mathbf{V}_j : \|\mathbf{V}_j\|_F^2 \leq P\}} \min_{\{\mathbf{\Delta}_{i,j} : \|\mathbf{\Delta}_{i,j}\|_F \leq \epsilon_{i,j}\}} \sum_{i=1}^{K} \sum_{l=1}^{d_i} w_i R_{i,l}. \tag{6.7}$$

Remark 6.2. *Here, we focus on the regime where quantization errors dominate the channel estimation errors. In this regime, the above formulation is appropriate. For channel estimation errors, this formulation can be applied by choosing the radii of error balls based on "allowable outages". However, we do not provide any details on this.*

Next, we proceed to obtain iterative algorithms for these two problems of interest.

6.4.1 Single-Stream Decoding with Perfect CSI

We start with the well-known relation between the achievable rate $R_{i,l}$ and the mean-square-error (MSE) with optimal receive filter. The MSE $e_{i,l}$ for the stream given in (6.3) is given by

$$e_{i,l} = \left|\mathbf{g}_{i,l}^{\dagger}\mathbf{H}_{i,i}\mathbf{v}_{i,l} - 1\right|^2 + \sum_{(j,k)\neq(i,l)} \left|\mathbf{g}_{i,l}^{\dagger}\mathbf{H}_{i,j}\mathbf{v}_{j,k}\right|^2 + \sigma^2 \left\|\mathbf{g}_{i,l}\right\|^2. \tag{6.8}$$

We denote the corresponding MSE with optimal receive filter by $\hat{e}_{i,l}$. The following lemma states the well-known relation.

Lemma 6.5. *The achievable rate $R_{i,l}$ in (6.4) and the MSE in (6.8) with optimal receive filter denoted by $\hat{e}_{i,l}$ have a one-to-one correspondence given by*

$$R_{i,l} = \log\left(\frac{1}{\hat{e}_{i,l}}\right), \forall i, l. \tag{6.9}$$

Next, in (6.5), first using Lemma 6.5 and then applying Lemma 6.3 with slack variables $\mathbf{s} = \{s_{i,l}\}$, we can reformulate the optimization problem

in (6.5) as follows:

$$\max_{\substack{\{\mathbf{V}_j:\|\mathbf{V}_j\|_F^2\leq P\},\\ \mathbf{s},\{\mathbf{G}_i\}}} \quad \sum_{i=1}^{K}\sum_{l=1}^{d_i} w_i(-e_{i,l}s_{i,l}+\log s_{i,l}+1), \tag{6.10}$$

where $e_{i,l}$ is the mean-square-error given by (6.8). This formulation has a weighted MSE minimization sub-problem that makes it more tractable.

Next, we show that the sub-problem in receive filters and slack variables can be solved optimally in closed-form.

Lemma 6.6. *Consider the sub-problem in (6.10) for any given* $\{\mathbf{V}_j\}, \forall j$. *Then, the optimal receive filters are given by*

$$\mathbf{g}_{i,l}^{opt} = \left(\sum_{j=1}^{K}\mathbf{H}_{i,j}\mathbf{V}_j\mathbf{V}_j^{\dagger}\mathbf{H}_{i,j}^{\dagger}+\sigma^2\mathbf{I}\right)^{-1}\mathbf{H}_{i,i}\mathbf{v}_{i,l}, \forall i,l, \tag{6.11}$$

which can be readily expressed in matrix form as

$$\mathbf{G}_{i}^{opt} = \left(\sum_{j=1}^{K}\mathbf{H}_{i,j}\mathbf{V}_j\mathbf{V}_j^{\dagger}\mathbf{H}_{i,j}^{\dagger}+\sigma^2\mathbf{I}\right)^{-1}\mathbf{H}_{i,i}\mathbf{V}_{i}, \forall i. \tag{6.12}$$

Furthermore, the optimal $\mathbf{s}$ *is given by*

$$s_{i,l}^{opt} = \frac{1}{e_{i,l}}, \forall i,l, \tag{6.13}$$

where $e_{i,l}$ *is obtained by substituting (6.11) in (6.8).*

Proof. The proof follows directly from standard results on optimal MSE receive filter [72] and Lemma 6.3. □

Now, for any given $\mathbf{s}$ and $\{\mathbf{G}_i\}$, the sub-problem of interest to solve (6.10) is the following minimum weighted MSE problem:

$$\min_{\{\mathbf{V}_j:\|\mathbf{V}_j\|_F^2\leq P\}} \sum_{i=1}^{K}\sum_{l=1}^{d_i} \alpha_{i,l} e_{i,l}, \tag{6.14}$$

where $\alpha_{i,l} = w_i s_{i,l}$. Substituting (6.8) in (6.14), the optimization problem in (6.14) becomes

$$\min_{\{\mathbf{V}_j:\|\mathbf{V}_j\|_F^2\leq P\}} \quad \sum_{i=1}^{K}\sum_{l=1}^{d_i} \alpha_{i,l}\Bigg(|\mathbf{g}_{i,l}^\dagger \mathbf{H}_{i,i}\mathbf{v}_{i,l} - 1|^2 + \sum_{(j,k)\neq(i,l)} |\mathbf{g}_{i,l}^\dagger \mathbf{H}_{i,j}\mathbf{v}_{j,k}|^2 + \sigma^2 \|\mathbf{g}_{i,l}\|^2 \Bigg). \tag{6.15}$$

Let $\mathbf{A}_i = \text{diag}\{\alpha_{i,1},\ldots,\alpha_{i,d_i}\}$. Then, the problem in (6.15) can be expressed in matrix form as

$$\min_{\{\mathbf{V}_j:\|\mathbf{V}_j\|_F^2\leq P\}} \quad \sum_{i=1}^{K}\Bigg(\left\|\mathbf{A}_i^{\frac{1}{2}}\mathbf{G}_i^\dagger \mathbf{H}_{i,i}\mathbf{V}_i - \mathbf{A}_i^{\frac{1}{2}}\right\|_F^2 + \sum_{j\neq i}\left\|\mathbf{A}_i^{\frac{1}{2}}\mathbf{G}_i^\dagger \mathbf{H}_{i,j}\mathbf{V}_j\right\|_F^2 + \sigma^2\left\|\mathbf{A}_i^{\frac{1}{2}}\mathbf{G}_i^\dagger\right\|_F^2 \Bigg). \tag{6.16}$$

Next, we show that this problem can be solved optimally in closed-form except for a scalar variable. Additionally, this scalar variable can be obtained efficiently using a bi-section search.

Lemma 6.7. *Consider the optimization problem in (6.16) for any given* $\mathbf{s}$ *and* $\{\mathbf{G}_i\}, \forall i$. *Then, the optimal transmit precoders are given by*

$$\mathbf{V}_j^{\text{opt}} = \left(\sum_{i=1}^{K} \mathbf{H}_{i,j}^\dagger \mathbf{G}_i \mathbf{A}_i \mathbf{G}_i^\dagger \mathbf{H}_{i,j} + \lambda_j \mathbf{I}\right)^{-1} \mathbf{H}_{j,j}^\dagger \mathbf{G}_j \mathbf{A}_j, \tag{6.17}$$

where each non-negative parameter λ_j is such that the power constraint

$$\left\|\mathbf{V}_j^{\text{opt}}\right\|_F^2 \leq P$$

is satisfied. If $\lambda_j > 0$, then this inequality should be satisfied with equality. Therefore, a unique λ_j exists.

Proof. Let

$$\mathbf{R}_j = \sum_{i=1}^{K} \mathbf{H}_{i,j}^{\dagger} \mathbf{G}_i \mathbf{A}_i \mathbf{G}_i^{\dagger} \mathbf{H}_{i,j} + \lambda_j \mathbf{I}.$$

The Lagrangian corresponding to this problem can be expressed as

$$\begin{aligned} \mathcal{L} \;=\; & \sum_{j=1}^{K} \Bigg(\left\| \mathbf{R}_j^{\frac{1}{2}} \mathbf{V}_j - \mathbf{R}_j^{-\frac{1}{2}} \mathbf{H}_{j,j}^{\dagger} \mathbf{G}_j \mathbf{A}_j \right\|_F^2 - \left\| \mathbf{R}_j^{-\frac{1}{2}} \mathbf{H}_{j,j}^{\dagger} \mathbf{G}_j \mathbf{A}_j \right\|_F^2 \\ & + \sigma^2 \left\| \mathbf{A}_j^{\frac{1}{2}} \mathbf{G}_j^{\dagger} \right\|_F^2 + \lambda_j P \Bigg). \end{aligned}$$

In fact, it is easier to verify that the above expression simplifies to the Lagrangian. Now, the proof follows from the fact that first term inside summation must be zero and complementary slackness conditions. □

The results in Lemma 6.6 and Lemma 6.7 lead to a natural (iterative) algorithm for weighted-sum rate maximization:

1. Initialize $\{\mathbf{V}_j\}$;
2. Update $\{\mathbf{G}_i\}, \mathbf{s}$ based on Lemma 6.6;
3. Update $\{\mathbf{V}_j\}$ based on Lemma 6.7;

4. Iterate above two steps till convergence or till maximum number of iterations is reached.

Theorem 6.8. *The above algorithm is guaranteed to converge.*

Proof. The updates in the algorithm result in monotone increase of objective function in (6.10). Since the objective function is bounded above, this monotonicity guarantees the convergence of this algorithm. □

We note that this result extends the one obtained in [28] for the MIMO broadcast channel under perfect CSI.

6.4.2 Single-Stream Decoding under Channel Uncertainty

In this section, we provide the additional steps required in dealing with channel uncertainty. By introducing slack variables $\mathbf{s} = \{s_{i,l}\}$, we can reformulate the optimization problem in (6.7) as follows:

$$\max_{\{\mathbf{V}_j:\|\mathbf{V}_j\|_F^2\leq P\}} \min_{\{\mathbf{\Delta}_{i,j}:\|\mathbf{\Delta}_{i,j}\|_F\leq\epsilon_{i,j}\}} \max_{\mathbf{s},\{\mathbf{G}_i\}} \sum_{i=1}^{K}\sum_{l=1}^{d_i} w_i(-e_{i,l}s_{i,l} + \log s_{i,l} + 1). \quad (6.18)$$

The steps used to arrive at this formulation is same as in the previous section, and hence omitted for brevity. However, this formulation does not directly lead to a useful algorithm. Therefore, we look at the max-min version of the inner min-max problem in (6.18) given by

$$\max_{\substack{\{\mathbf{V}_j:\|\mathbf{V}_j\|_F^2\leq P\},\\ \mathbf{s},\{\mathbf{G}_i\}}} \min_{\{\mathbf{\Delta}_{i,j}:\|\mathbf{\Delta}_{i,j}\|_F\leq\epsilon_{i,j}\}} \sum_{i=1}^{K}\sum_{l=1}^{d_i} w_i(-e_{i,l}s_{i,l} + \log s_{i,l} + 1). \quad (6.19)$$

This formulation is relevant due to the following two reasons: (i) It is an achievable weighted-sum rate; (ii) In the algorithm, none of the optimization variables should depend on the perfect channel.

Next, we show that three sub-problems corresponding to the outer maximization in (6.19) can be solved optimally, leading to an iterative algorithm. Towards showing this, the key step is the derivation of an equivalent problem for the inner minimization in (6.19), which is given in next lemma.

Lemma 6.9. *The inner minimization in (6.19) is equivalent (in terms of objective) to the following:*

$$\begin{aligned} \max_{\{\tau_{i,j},\lambda_{i,j}\}} \quad & \sum_{i=1}^{K} w_i \left(-\sum_{j=1}^{K} \tau_{i,j} - \sigma^2 \left\| \mathbf{G}_i \mathbf{B}_i \right\|_F^2 + 2\log\left|\mathbf{B}_i\right| + d_i \right) \qquad (6.20) \\ s.t. \quad & \begin{bmatrix} \tau_{i,j} - \lambda_{i,j} & \mathbf{c}_{i,j}^{\dagger} & \mathbf{0} \\ \mathbf{c}_{i,j} & \mathbf{I} & -\epsilon_{i,j}\mathbf{C}_{i,j} \\ \mathbf{0} & -\epsilon_{i,j}\mathbf{C}_{i,j}^{\dagger} & \lambda_{i,j}\mathbf{I} \end{bmatrix} \succeq 0, \forall i,j, \\ & \lambda_{i,j} \geq 0, \forall i,j. \end{aligned}$$

Proof. For this derivation, we introduce the following notation:

$$\mathbf{B}_i = \mathrm{diag}\{\sqrt{s_{i,1}}, \ldots, \sqrt{s_{i,d_i}}\},$$

the MSE matrix defined as

$$\begin{aligned} \mathbf{E}_i &= \mathbb{E}\left[\left(\hat{\mathbf{y}}_i - \hat{\mathbf{x}}_j \right) \left(\hat{\mathbf{y}}_i - \hat{\mathbf{x}}_j \right)^{\dagger} \right], \\ &= \left(\mathbf{G}_i^{\dagger}\mathbf{H}_{i,i}\mathbf{V}_i - \mathbf{I} \right) \left(\mathbf{G}_i^{\dagger}\mathbf{H}_{i,i}\mathbf{V}_i - \mathbf{I} \right)^{\dagger} + \sum_{j \neq i} \mathbf{G}_i^{\dagger}\mathbf{H}_{i,j}\mathbf{V}_j\mathbf{V}_j^{\dagger}\mathbf{H}_{i,j}^{\dagger}\mathbf{G}_i \\ &\quad + \sigma^2 \mathbf{G}_i^{\dagger}\mathbf{G}_i, \end{aligned}$$

and

$$\mathbf{T}_{i,j} = \left(\mathbf{G}_i^\dagger \mathbf{H}_{i,j} \mathbf{V}_j - \delta_{i,j}\mathbf{I}\right)\left(\mathbf{G}_i^\dagger \mathbf{H}_{i,j} \mathbf{V}_j - \delta_{i,j}\mathbf{I}\right)^\dagger.$$

Now, using the above notation, the inner minimization in (6.19) can be expressed in matrix form:

$$\min_{\{\mathbf{\Delta}_{i,j}:\|\mathbf{\Delta}_{i,j}\|_F \le \epsilon_{i,j}\}} \sum_{i=1}^{K} w_i \left(-\operatorname{tr}\left(\mathbf{E}_i \mathbf{B}_i \mathbf{B}_i^\dagger\right) + 2\log|\mathbf{B}_i| + d_i\right). \tag{6.21}$$

Due to separation of variables, we can focus on the following optimization problems: for all i and j

$$\min_{\mathbf{\Delta}_{i,j}:\|\mathbf{\Delta}_{i,j}\|_F \le \epsilon_{i,j}} -\operatorname{tr}\left(\mathbf{T}_{i,j}\mathbf{B}_i\mathbf{B}_i^\dagger\right). \tag{6.22}$$

For each i and j, by introducing a slack variable $\tau_{i,j}$, we obtain the following problem with same optimal value as (6.22):

$$\begin{aligned} \max_{\tau_{i,j}} \quad & -\tau_{i,j} && (6.23)\\ \text{s.t.} \quad & \left\|\mathbf{B}_i^\dagger\left(\mathbf{G}_i^\dagger\left(\hat{\mathbf{H}}_{i,j} + \mathbf{\Delta}_{i,j}\right)\mathbf{V}_j - \delta_{i,j}\mathbf{I}\right)\right\|_F^2 \le \tau_{i,j}, \\ & \forall\left\{\mathbf{\Delta}_{i,j} : \|\mathbf{\Delta}_{i,j}\|_F \le \epsilon_{i,j}\right\}. && (6.24) \end{aligned}$$

To proceed further, we introduce the following notation:

$$\begin{aligned} \mathbf{c}_{i,j} &= \operatorname{vec}\left(\mathbf{B}_i^\dagger\left(\mathbf{G}_i^\dagger\hat{\mathbf{H}}_{i,j}\mathbf{V}_j - \delta_{i,j}\mathbf{I}\right)\right), \\ \mathbf{C}_{i,j} &= \mathbf{V}_j^\dagger \otimes \left(\mathbf{B}_i^\dagger\mathbf{G}_i^\dagger\right), \\ \mathbf{d}_{i,j} &= \operatorname{vec}\left(\mathbf{\Delta}_{i,j}\right). \end{aligned}$$

Here, $\otimes$ denotes Kronecker product. Now, using the fact

$$\operatorname{vec}(\mathbf{ABC}) = \left(\mathbf{C}^\dagger \otimes \mathbf{A}\right)\operatorname{vec}(\mathbf{B}),$$

we can express (6.24) as

$$\|\mathbf{c}_{i,j} + \mathbf{C}_{i,j}\mathbf{d}_{i,j}\|_2^2 \leq \tau_{i,j}, \forall \left\{\mathbf{d}_{i,j} : \|\mathbf{d}_{i,j}\|_2 \leq \epsilon_{i,j}\right\}. \tag{6.25}$$

Using Schur complement lemma (Lemma 6.1), (6.25) is equivalent to

$$\begin{bmatrix} \tau_{i,j} & \mathbf{c}_{i,j}^\dagger \\ \mathbf{c}_{i,j} & \mathbf{I} \end{bmatrix} + \begin{bmatrix} 0 & \mathbf{d}_{i,j}^\dagger \mathbf{C}_{i,j}^\dagger \\ \mathbf{C}_{i,j}\mathbf{d}_{i,j} & \mathbf{0} \end{bmatrix} \succeq 0, \forall \left\{\mathbf{d}_{i,j} : \|\mathbf{d}_{i,j}\|_2 \leq \epsilon_{i,j}\right\}. \tag{6.26}$$

Now, applying Lemma 6.2, (6.26) is equivalent to

$$\exists \lambda_{i,j} \geq 0, \begin{bmatrix} \tau_{i,j} - \lambda_{i,j} & \mathbf{c}_{i,j}^\dagger & \mathbf{0} \\ \mathbf{c}_{i,j} & \mathbf{I} & -\epsilon_{i,j}\mathbf{C}_{i,j} \\ \mathbf{0} & -\epsilon_{i,j}\mathbf{C}_{i,j}^\dagger & \lambda_{i,j}\mathbf{I} \end{bmatrix} \succeq 0. \tag{6.27}$$

This completes the proof. □

Next, we show that when each set of the variables $\{\mathbf{B}_i\}$, $\{\mathbf{G}_i\}$ and $\{\mathbf{V}_j\}$ can be solved optimally while the remaining two sets are fixed. From Lemma 6.9, the optimization problem in (6.19) can be expressed as

$$\begin{aligned} \max_{\substack{\{\mathbf{V}_j : \|\mathbf{V}_j\|_F^2 \leq P\}, \\ \{\mathbf{B}_i\}, \{\mathbf{G}_i\}}} \quad \max_{\{\tau_{i,j}, \lambda_{i,j}\}} \quad & \sum_{i=1}^{K} w_i \Bigg(- \sum_{j=1}^{K} \tau_{i,j} - \sigma^2 \|\mathbf{G}_i \mathbf{B}_i\|_F^2 \\ & \qquad + 2\log|\mathbf{B}_i| + d_i \Bigg) \\ \text{s.t.} \quad & \begin{bmatrix} \tau_{i,j} - \lambda_{i,j} & \mathbf{c}_{i,j}^\dagger & \mathbf{0} \\ \mathbf{c}_{i,j} & \mathbf{I} & -\epsilon_{i,j}\mathbf{C}_{i,j} \\ \mathbf{0} & -\epsilon_{i,j}\mathbf{C}_{i,j}^\dagger & \lambda_{i,j}\mathbf{I} \end{bmatrix} \succeq 0, \forall i, j, \\ & \lambda_{i,j} \geq 0, \forall i, j. \end{aligned} \tag{6.28}$$

Now, note that the constraints of (6.28) are linear matrix inequalities in each set of variables while others two sets are fixed. This immediately leads to the following results.

Lemma 6.10. *Consider the sub-problem in (6.28) for any given* $\{\mathbf{V}_j\}$ *and* $\{\mathbf{G}_i\}$*. Then, the optimization for each* i *separates, and each* $\mathbf{B}_i$ *can be solved efficiently by solving the resulting MAX-DET in (6.28) along with the constraint that* $\mathbf{B}_i$ *is diagonal.*

Lemma 6.11. *Consider the sub-problem in (6.28) for any given* $\{\mathbf{V}_j\}$ *and* $\{\mathbf{B}_i\}$*. Then, the optimization for each* i *separates, and each* $\mathbf{G}_i$ *can be solved efficiently by solving the resulting SDP in (6.28). Similarly, consider the sub-problem in (6.28) for any given* $\{\mathbf{B}_i\}$ *and* $\{\mathbf{G}_i\}$*. Then, the optimization for each* j *separates, and* $\mathbf{V}_j$ *can be solved efficiently by solving the resulting SDP.*

The above results lead to an iterative algorithm for obtaining transmit precoders, with the following guarantee.

Theorem 6.12. *This iterative algorithm is monotone in the objective function in (6.28).*

We remark that it often difficult to provide a monotone algorithm for a max-min optimization problem such as (6.19). As a consequence of this monotonicity property, this algorithm is provably convergent.

6.5 Precoder Design with Single-User Decoding

In this section, we focus on the communication scheme that uses single-user decoding. Since there are many similarities with the previous section, we just describe the additional steps. With single-user decoding, the interference

from other transmitters are treated as noise. We denote the effective transmit symbols at j-th transmitter by $\hat{\mathbf{x}}_j \in \mathbb{C}^{d_j \times 1}$ and the precoder at j-th transmitter by $\mathbf{V}_j \in \mathbb{C}^{M_j \times d_j}$. Then, the received signal at the i-th receiver is given by

$$\mathbf{y}_i = \sum_{j=1}^{K} \mathbf{H}_{i,j}\mathbf{V}_j\hat{\mathbf{x}}_j + \mathbf{z}_i. \tag{6.29}$$

Now, using single-user decoding, the rate achievable by this receiver is

$$R_i = \log \left| \mathbf{I} + \mathbf{V}_i^{\dagger}\mathbf{H}_{i,i}^{\dagger} \left(\sum_{j \neq i} \mathbf{H}_{i,j}\mathbf{V}_j\mathbf{V}_j^{\dagger}\mathbf{H}_{i,j}^{\dagger} + \sigma^2\mathbf{I} \right)^{-1} \mathbf{H}_{i,i}\mathbf{V}_i \right|. \tag{6.30}$$

We consider the problem of obtaining the linear transmit precoders that maximize the weighted sum-rate, which can be formally stated as:

$$\begin{aligned} \max_{\{\mathbf{V}_j\}} \quad & \sum_{i=1}^{K} w_i R_i \\ \text{s.t.} \quad & \|\mathbf{V}_j\|_F^2 \leq P, \forall j, \end{aligned} \tag{6.31}$$

where $w_i \in \mathbb{R}_+$ are given weights and R_i in (6.31) is given by (6.30). Furthermore, we consider a robust counterpart of this problem in the presence of bounded channel errors:

$$\max_{\{\mathbf{V}_j : \|\mathbf{V}_j\|_F^2 \leq P\}} \min_{\{\mathbf{\Delta}_{i,j} : \|\mathbf{\Delta}_{i,j}\|_F \leq \epsilon_{i,j}\}} \sum_{i=1}^{K} w_i R_i. \tag{6.32}$$

Recall that the MSE matrix of the received signal in (6.29) with receive filter $\mathbf{G}_i \in \mathbb{C}^{N_i \times d_i}$ is defined as

$$\mathbf{E}_i = \mathbb{E}\left[\left(\mathbf{G}_i^{\dagger}\mathbf{y}_i - \hat{\mathbf{x}}_j \right) \left(\mathbf{G}_i^{\dagger}\mathbf{y}_i - \hat{\mathbf{x}}_j \right)^{\dagger} \right]. \tag{6.33}$$

Let $\hat{\mathbf{G}}_i$ denote the optimal receive filter, i.e.,

$$\hat{\mathbf{G}}_i = \arg\min_{\mathbf{G}_i} \operatorname{tr}(\mathbf{E}_i), \tag{6.34}$$

and let $\hat{\mathbf{E}}_i$ be the MSE matrix with the receive filter in (6.34) substituted in (6.33). For given transmit precoders, the optimal receive filters are same as in the previous section and is given by (6.12). Note that the optimal filter does not change even if the objective is changed to $\operatorname{tr}(\mathbf{S}\mathbf{E}_i)$ for any given $\mathbf{S} \succ 0$.

Next, we provide a lemma that relates the determinant of optimal MSE matrix to the achievable rate.

Lemma 6.13. *The achievable rate R_i in (6.30) and the MSE matrix with optimal receive filters $\hat{\mathbf{E}}_i$ are related as follows:*

$$R_i = \log \left| \hat{\mathbf{E}}_i^{-1} \right|, \forall i. \tag{6.35}$$

Now, using Lemma 6.13 and applying Lemma 6.4 with (matrix) slack variables $\mathbf{S}_i \in \mathbb{C}^{d_i \times d_i}$, we reformulate the optimization problem in (6.31) as follows:

$$\max_{\substack{\{\mathbf{V}_j : \|\mathbf{V}_j\|_F^2 \leq P\}, \\ \{\mathbf{G}_i\}, \{\mathbf{S}_i\}}} \quad \sum_{i=1}^{K} w_i \left(-\operatorname{tr}(\mathbf{S}_i \mathbf{E}_i) + \log |\mathbf{S}_i| + d_i \right). \tag{6.36}$$

With the formulation in (6.36), we can apply the results in the previous section. For any given $\{\mathbf{V}_j\}$, the optimal $\{\mathbf{G}_i\}$ and $\{\mathbf{S}_i\}$ are given by (6.12) and $\mathbf{S}_i = \mathbf{E}_i^{-1}$, respectively. For any given $\{\mathbf{G}_i\}$ and $\{\mathbf{S}_i\}$, optimal $\mathbf{V}_j$ is given by (6.17) with $\mathbf{A}_i$ replaced by $w_i \mathbf{S}_i$. Now, it is straightforward to obtain provably convergent iterative algorithms for obtaining the transmit precoder with perfect CSI and under channel uncertainty.

6.6 Robust Max-Min Rate Objective

So far, we dealt with the objective of weighted-sum rate. This objective is particularly applicable when a higher-level scheduler ensures long-term fairness (or, in general, any network optimization) by adapting the weights assigned to users with time. However, in certain scenarios, we prefer short-term fairness. An objective that captures short-term fairness is the max-min rate,[4] which implies maximization of the minimum achievable rate in the system. Note that neither of the objectives (weighted-sum or max-min) include the other as special case.

Motivated from short-term fairness, next, we study the precoder design problem that arise in this context. Consider single-user decoding that leads to the set of achievable rates $\{R_i\}$ in (6.30). A natural robust precoder design problem in the presence of bounded errors is:

$$\max_{\{\mathbf{V}_j:\|\mathbf{V}_j\|_F^2\leq P\}} \min_{\{\mathbf{\Delta}_{i,j}:\|\mathbf{\Delta}_{i,j}\|_F\leq\epsilon_{i,j}\}} \min_i R_i. \tag{6.37}$$

For this problem, finding the optimal solution does not seem tractable.[5] Therefore, we apply the same methodology as before - solve sub-problems optimally to obtain an iterative algorithm.

We follow the same steps as in the previous section to arrive at the

[4]The results will immediately generalize to max-min weighted-rate, which can be used to incorporate transmitter/receiver priorities.

[5]Note that this problem is tractable in the MISO setting.

following equivalent problem:

$$\max_{\{\mathbf{V}_j:\|\mathbf{V}_j\|_F^2\le P\}}\ \min_{\{\mathbf{\Delta}_{i,j}:\|\mathbf{\Delta}_{i,j}\|_F\le\epsilon_{i,j}\}}\ \min_i\ \max_{\{\mathbf{G}_i\},\{\mathbf{S}_i\}}\ \left(-\operatorname{tr}(\mathbf{S}_i\mathbf{E}_i)+\log|\mathbf{S}_i|+d_i\right). \quad (6.38)$$

Next, to keep the problem tractable, we consider an achievable lower bound obtained by switching min-max to max-min:

$$\max_{\substack{\{\mathbf{V}_j:\|\mathbf{V}_j\|_F^2\le P\},\\ \{\mathbf{G}_i\},\{\mathbf{S}_i\}}}\ \min_{\{\mathbf{\Delta}_{i,j}:\|\mathbf{\Delta}_{i,j}\|_F\le\epsilon_{i,j}\}}\ \min_i\left(-\operatorname{tr}(\mathbf{S}_i\mathbf{E}_i)+\log|\mathbf{S}_i|+d_i\right). \quad (6.39)$$

Now, by introducing a slack variable β and following the steps in Section 6.4.2, we can reformulate the problem in (6.39) as

$$\begin{aligned}\max_{\substack{\{\mathbf{V}_j:\|\mathbf{V}_j\|_F^2\le P\},\\ \{\mathbf{G}_i\},\{\mathbf{B}_i\}}}\ \max_{\{\tau_{i,j},\lambda_{i,j},\beta\}}\quad & \beta \quad (6.40)\\ \text{s.t.}\quad & -\sum_{j=1}^{K}\tau_{i,j}-\sigma^2\left\|\mathbf{G}_i\mathbf{B}_i\right\|_F^2+2\log|\mathbf{B}_i|+d_i\ge\beta,\forall i,\\ & \begin{bmatrix}\tau_{i,j}-\lambda_{i,j} & \mathbf{c}_{i,j}^\dagger & \mathbf{0}\\ \mathbf{c}_{i,j} & \mathbf{I} & -\epsilon_{i,j}\mathbf{C}_{i,j}\\ \mathbf{0} & -\epsilon_{i,j}\mathbf{C}_{i,j}^\dagger & \lambda_{i,j}\mathbf{I}\end{bmatrix}\succeq 0,\forall i,j,\\ & \lambda_{i,j}\ge 0,\forall i,j.\end{aligned}$$

For the outer maximization in (6.39), individual maximizations in each set of variables can be solved optimally while the other two set of variables are fixed. Particularly, for the individual maximization in $\{\mathbf{B}_i\}$, the problems separate into K MAX-DETs (same as in Section 6.4.2). Note that the variable β does not have any explicit role as the problems separate. Similarly, the optimization in $\{\mathbf{G}_i\}$ also separates into K SDPs (same as in Section 6.4.2). However, the difference is the optimization in the precoder variables. For this

optimization in $\{\mathbf{V}_i\}$, we need to solve one joint SDP that results from (6.39) while $\{\mathbf{B}_i\}$ and $\{\mathbf{G}_i\}$ are fixed. Thus, the above-mentioned three optimizations provide an iterative algorithm for robust precoder design under max-min objective, which is guaranteed to converge.

6.7 Genie-MAC Outer Bound

An optimization framework based on *genie*-MACs for obtaining outer bounds for K-user SISO GIFCs is given in Section 5.3. Even though the optimization framework is for general GIFCs, its evaluation for MIMO GIFCs has not been addressed so far. Next, we generalize the genie-MAC outer bounding technique to incorporate channel uncertainty using notions of compound-MAC capacity and then obtain computable outer bounds using an alternating optimization approach, where each of the sub-problems are formulated as convex MAX-DET problems. Interestingly, this is one of the first approaches to obtain tighter outer bounds for the GIFC in the presence of channel uncertainty.

6.7.1 Outer Bound on Capacity Region with Perfect CSI

Consider any permutation function $\pi : \{1, 2, \ldots, K\} \mapsto \{1, 2, \ldots, K\}$, and integers k such that $1 \leq k \leq K$. For each $j \in \{1, \ldots, k\}$, we define

$$\mathbf{U}_j = \begin{bmatrix} \mathbf{H}_{\pi(1),\pi(j)} \\ \vdots \\ \mathbf{H}_{\pi(j),\pi(j)} \end{bmatrix},$$

and introduce variables $\mathbf{L}_j \in \mathbb{C}^{\left(\sum_{i=j+1}^{k} N_{\pi(i)}\right) \times M_{\pi(j)}}$. Let

$$\mathbf{F}_j = \begin{bmatrix} \mathbf{U}_j \\ \mathbf{L}_j \end{bmatrix}.$$

Now, by extending the optimization framework in Section 5.3 to MIMO GIFCs, an outer bound on $\sum_{j=1}^{k} R_{\pi(j)}$ is given by the following:

$$\begin{aligned} \min_{\mathbf{\Sigma} \in \mathcal{T}, \{\mathbf{L}_j\}} \quad \max_{\{\mathbf{S}_j\}} \quad & \log \left| \mathbf{\Sigma} + \sum_{j=1}^{k} \mathbf{F}_j \mathbf{S}_j \mathbf{F}_j^{\dagger} \right| - \log |\mathbf{\Sigma}| \\ \text{s.t.} \quad & \operatorname{tr}(\mathbf{S}_j) \leq P, \forall j, \\ & \mathbf{S}_j \succeq 0, \forall j, \end{aligned} \tag{6.41}$$

where

$$\mathcal{T} = \left\{ \mathbf{T} | \mathbf{T} = [\mathbf{T}_{i,l}]_{i,l}, \mathbf{T}_{i,l} \in \mathbb{C}^{N_{\pi(i)} \times N_{\pi(l)}}, \mathbf{T}_{i,i} = \sigma^2 \mathbf{I}, \mathbf{T}_{l,i} = \mathbf{T}_{i,l}^{\dagger}, \forall i, l, \mathbf{T} \succ \mathbf{0} \right\}.$$

The inner maximization problem in (6.41) is convex, which is well-studied in MIMO-MAC literature. As shown in [138], a dual problem can be derived for this problem whose optimum value matches the primal problem. Using this dual, the problem in (6.41) can be reformulated as:

$$\begin{aligned} \min_{\substack{\mathbf{\Sigma} \in \mathcal{T}, \{\mathbf{L}_j\}, \\ \mathbf{\Gamma}, \{\lambda_j\}}} \quad & -\log |\mathbf{\Gamma}| - \log |\mathbf{\Sigma}| + \operatorname{tr}(\mathbf{\Gamma}\mathbf{\Sigma}) + \sum_{j=1}^{k} \lambda_j P - \sum_{i=1}^{k} N_{\pi(i)} \\ \text{s.t.} \quad & \lambda_j \mathbf{I} \succeq \mathbf{F}_j^{\dagger} \mathbf{\Gamma} \mathbf{F}_j, \forall j, \\ & \mathbf{\Gamma} \succ 0. \end{aligned} \tag{6.42}$$

Now, the following lemmas are immediate.

Lemma 6.14. *Consider any given $\mathbf{\Gamma} \succ 0$. Then, the optimization in (6.42) over remaining variables, given by*

$$\begin{aligned} \min_{\mathbf{\Sigma},\{\mathbf{L}_j\},\{\lambda_j\}} \quad & -\log|\mathbf{\Gamma}| - \log|\mathbf{\Sigma}| + \mathrm{tr}\,(\mathbf{\Gamma}\mathbf{\Sigma}) + \sum_{j=1}^{k} \lambda_j P - \sum_{i=1}^{k} N_{\pi(i)} \quad (6.43) \\ s.t. \quad & \left[\begin{array}{cc} \lambda_j \mathbf{I} & \mathbf{F}_j^* \\ \mathbf{F}_j & \mathbf{\Gamma}^{-1} \end{array} \right] \succeq 0, \forall j, \\ & \mathbf{\Sigma} \in \mathcal{T}, \end{aligned}$$

is a MAX-DET problem that can be solved efficiently.

Proof. For any given $\mathbf{\Gamma} \succ 0$, by Shur complement lemma, $\lambda_j \mathbf{I} \succeq \mathbf{F}_j^\dagger \mathbf{\Gamma} \mathbf{F}_j$ is equivalent to

$$\left[\begin{array}{cc} \lambda_j \mathbf{I} & \mathbf{F}_j^* \\ \mathbf{F}_j & \mathbf{\Gamma}^{-1} \end{array} \right] \succeq 0.$$

Now, the proof follows. □

Lemma 6.15. *Consider any given $\mathbf{\Sigma} \in \mathcal{T}$ and $\{\mathbf{L}_j\}$. Then, the optimization in (6.42) over remaining variables, given by*

$$\begin{aligned} \min_{\mathbf{\Gamma},\{\lambda_j\}} \quad & -\log|\mathbf{\Gamma}| - \log|\mathbf{\Sigma}| + \mathrm{tr}\,(\mathbf{\Gamma}\mathbf{\Sigma}) + \sum_{j=1}^{k} \lambda_j P - \sum_{i=1}^{k} N_{\pi(i)} \quad (6.44) \\ s.t. \quad & \lambda_j \mathbf{I} \succeq \mathbf{F}_j^* \mathbf{\Gamma} \mathbf{F}_j, \forall j, \\ & \mathbf{\Gamma} \succ 0, \end{aligned}$$

is a MAX-DET problem that can be solved efficiently.

Note that both optimization problems in (6.43) and (6.44) are convex problems. These can be solved using interior-point methods developed for determinant maximization with linear matrix inequality constraints. The problems in (6.43) and (6.44) lead to an iterative algorithm that converges.

6.7.2 Outer Bound on Capacity Region under Channel Uncertainty

Consider any K user MIMO GIFC. Let n denote the number of channel uses (or the codeword length). For all $1 \leq j \leq K$, let $R_j(n)$ denote the rate of the codebook $\mathcal{C}_j(n)$ (in bits per channel use) used by j-thtransmitter. Each $\mathcal{C}_j(n)$ satisfies the power constraint at the transmitter. For each n, we refer to $\underline{\mathcal{C}}(n) = (\mathcal{C}_1(n), \ldots, \mathcal{C}_K(n))$ as a *multi-user code.*

Now, consider a sequence of multi-user codes $\{\underline{\mathcal{C}}(n)\}_{n=1}^{\infty}$ with

$$\liminf_n R_j(n) = R_j, \forall 1 \leq j \leq K.$$

For any given set of channel estimates $\{\hat{H}_{i,j}\}$ (available at all transmitters) and codes $\{\mathcal{C}_j(n)\}$, let $\epsilon_i(n, \{\mathbf{\Delta}_{i,j}\})$ denote the error probability at i-th receiver obtained with some decoding scheme (with perfect CSIR) when the perturbations are $\{\mathbf{\Delta}_{i,j}\}$. Note that the average is over the space of codeword and noise realizations. Define

$$\epsilon_i(n) = \sup_{\{\mathbf{\Delta}_{i,j}\} \in \mathbf{\Delta}} \epsilon_i(n, \{\mathbf{\Delta}_{i,j}\}), \forall i.$$

Then the *rates* $(R_1, \ldots, R_K)$ *are defined to be achievable in the GIFC (under channel uncertainty) if the given sequence of multi-user codes satisfies*

$$\limsup_n \epsilon_i(n) = 0, \forall i. \tag{6.45}$$

Suppose that we are given a sequence of multi-user codes whose corresponding rate-tuple $(R_1, \ldots, R_K)$ is achievable in the GIFC (under channel uncertainty). Now consider the genie-MACs. For simplicity of description,

we consider those genie-MACs with all transmitters and identity permutation function. However, the arguments immediately follow for all cases. Let the channel estimates (available at all transmitters) be $\{\hat{H}_{i,j}\}$. Suppose that we use the multi-user code $\{\mathcal{C}_j(n)\}_{j=1}^{K}$ for some codeword length n. Assume perfect CSI at the MAC receiver which uses SIC in the order $1 \leq i \leq K$. We will use the fact that the joint error probability yielded by the SIC receiver is identical to that yielded by a genie assisted SIC receiver in which the genie always ensures perfect cancellation. Thus, the joint error probability at the MAC receiver for any given perturbations $\{\mathbf{\Delta}_{i,j}\}$ can be bounded as

$$\Pr(E_n(\{\mathbf{\Delta}_{i,j}\})) \leq \sum_{i=1}^{K} \Pr(E_{n,i}^{g}(\{\mathbf{\Delta}_{i,j}\})) \tag{6.46}$$

where $\Pr(E_{n,i}^{g}(\{\mathbf{\Delta}_{i,j}\}))$ is the codeword error probability of the codeword transmitted by the i-th transmitter obtained at another MAC receiver, say RX_i^g, which also uses SIC but for which signals corresponding to transmitters $1 \leq j \leq i-1$ are expurgated a-priori. Note that

$$\Pr(E_{n,i}^{g}(\{\mathbf{\Delta}_{i,j}\})) \leq \epsilon_i(n, \{\mathbf{\Delta}_{i,j}\}). \tag{6.47}$$

This follows from the fact that even a subset of observations available to receiver RX_i^g are statistically better than the ones available to the i-th receiver in the original GIFC. Thus, since K is finite, from (6.47), (6.46) and (6.45), we have

$$\limsup_{n} \sup_{\{\mathbf{\Delta}_{i,j}\} \in \mathbf{\Delta}} \Pr(E_n(\{\mathbf{\Delta}_{i,j}\})) = 0. \tag{6.48}$$

It can be shown that (6.47) and hence (6.48) is satisfied for a genie MAC with any noise correlation matrix $\mathbf{\Sigma} \succ \mathbf{0} : \mathbf{\Sigma}_{i,i} = \sigma^2\mathbf{I},\ \forall\ i$ and for arbitrarily fixed matrices $\mathbf{H}_{i,j}\ \forall\ i > j$. This observation leads to the following theorem.

Theorem 6.16. *$(R_1, \ldots, R_K)$ lies in the compound capacity region of a MAC with perfect CSIR given by*

$$\bigcup_{\mathbf{S}_j \succ \mathbf{0}, \mathrm{tr}(\mathbf{S}_j) \leq P,\ \forall\ j} \mathcal{R}(\{\mathbf{S}_j\}, \{\hat{H}_{i,j}\}). \tag{6.49}$$

The region $\mathcal{R}(\{\mathbf{S}_j\}, \{\hat{H}_{i,j}\})$ is given by

$$\left\{ \mathbf{r} : \sum_{j \in \mathcal{S}} r_j \leq \inf_{\substack{\tilde{\pi}; \mathbf{L}_{\tilde{\pi},j} \in \mathcal{L}_{\tilde{\pi},j}; \\ \{\mathbf{\Delta}_{i,j}\}; \mathbf{\Sigma}_{\tilde{\pi}} \in \mathcal{T}_{\tilde{\pi}}}} \log \left| \mathbf{I} + \mathbf{\Sigma}_{\tilde{\pi}}^{-1} \sum_{j \in \mathcal{S}} \mathbf{F}_{\tilde{\pi},j} \mathbf{S}_j \mathbf{F}_{\tilde{\pi},j}^{\dagger} \right|, \forall \mathcal{S} \right\},$$

where $\tilde{\pi} : \{1, \cdots, |\mathcal{S}|\} \to \mathcal{S}$ is a permutation, $\mathcal{L}_{\tilde{\pi},j} \subset \mathbb{C}^{\sum_{k=\tilde{\pi}^{-1}(j)+1}^{|\mathcal{S}|} N_{\tilde{\pi}(k)} \times M_j}$ is an arbitrarily large albeit compact set,

$$\mathbf{F}_{\tilde{\pi},j} = \begin{bmatrix} \hat{\mathbf{H}}_{\tilde{\pi}(1),j} + \mathbf{\Delta}_{\tilde{\pi}(1),j} \\ \vdots \\ \hat{\mathbf{H}}_{\tilde{\pi}(\tilde{\pi}^{-1}(j)-1),j} + \mathbf{\Delta}_{\tilde{\pi}(\tilde{\pi}^{-1}(j)-1),j} \\ \hat{\mathbf{H}}_{j,j} + \mathbf{\Delta}_{j,j} \\ \mathbf{L}_{\tilde{\pi},j} \end{bmatrix}, \forall j \in \mathcal{S}$$

and

$$\mathcal{T}_{\tilde{\pi}} = \Big\{ \mathbf{T} | \mathbf{T} = [\mathbf{T}_{i,l}]_{i,l}, \mathbf{T}_{i,l} \in \mathbb{C}^{N_{\tilde{\pi}(i)} \times N_{\tilde{\pi}(l)}}; \mathbf{T}_{i,i} = \sigma^2\mathbf{I}, \\ \mathbf{T}_{l,i} = \mathbf{T}_{i,l}^{\dagger}\ \forall i, l \in \{1, \ldots, |\mathcal{S}|\}; \mathbf{T} \succ \mathbf{0} \Big\}.$$

For an upper bound on robust weighted-sum rate, we can consider the compound MAC capacity region in (6.49). Note that we can consider

the subsets $\mathcal{S} \subseteq \{1, \ldots, K\}$ independently, i.e., take separate infimum. To illustrate the computation of an outer bound, we consider $\mathcal{S} = \{1, \cdots, k\}$ for some $1 \leq k \leq K$ and fix a permutation $\tilde{\pi}(\cdot)$ such that $\tilde{\pi}(j) = j \ \forall \ j \in \{1, \cdots, k\}$. Then, an upper bound on the sum of the robust (worse-case) rates $\sum_{j=1}^{k} R_j$ is given by the following (where we drop the subscript $\tilde{\pi}$):

$$\max_{\{\mathbf{S}_j\}} \min_{\substack{\{\boldsymbol{\Delta}_{i,j}: \|\boldsymbol{\Delta}_{i,j}\|_F \leq \epsilon_{i,j}\}, \\ \boldsymbol{\Sigma} \in \mathcal{T}, \{\mathbf{L}_j\}}} \log \left| \boldsymbol{\Sigma} + \sum_{j=1}^{k} \mathbf{F}_j \mathbf{S}_j \mathbf{F}_j^* \right| - \log |\boldsymbol{\Sigma}|$$

subject to $\mathrm{tr}\,(\mathbf{S}_j) \leq P, \forall j, \mathbf{S}_j \succeq 0, \forall j$. Now, we can obtain an upper bound on this max-min problem by considering the corresponding min-max problem:

$$\begin{aligned} \min_{\{\boldsymbol{\Delta}_{i,j}\}} \min_{\boldsymbol{\Sigma}, \{\mathbf{L}_j\}} \max_{\{\mathbf{S}_j\}} \ & \log \left| \boldsymbol{\Sigma} + \sum_{j=1}^{k} \mathbf{F}_j \mathbf{S}_j \mathbf{F}_j^* \right| - \log |\boldsymbol{\Sigma}| \\ \text{s.t.} \quad & \mathrm{tr}\,(\mathbf{S}_j) \leq P; \mathbf{S}_j \succeq 0, \forall j. \end{aligned} \tag{6.50}$$

Using the dual formulation for the inner maximization given in (6.50), we can rewrite (6.50) as

$$\begin{aligned} \min_{\substack{\{\boldsymbol{\Delta}_{i,j}\}, \\ \boldsymbol{\Sigma}, \{\mathbf{L}_j\}, \boldsymbol{\Gamma}, \{\lambda_j\}}} \ & -\log |\boldsymbol{\Gamma}| - \log |\boldsymbol{\Sigma}| + \mathrm{tr}\,(\boldsymbol{\Gamma}\boldsymbol{\Sigma}) + \sum_{j=1}^{k} \lambda_j P - \sum_{i=1}^{k} N_i \\ \text{s.t.} \quad & \lambda_j \mathbf{I} \succeq \mathbf{F}_j^* \boldsymbol{\Gamma} \mathbf{F}_j, \forall j; \boldsymbol{\Gamma} \succ 0. \end{aligned} \tag{6.51}$$

For fixed $\boldsymbol{\Gamma}$, we use the Shur complement lemma to obtain the following MAX-DET formulation in the other variables:

$$\begin{aligned} \min_{\substack{\{\boldsymbol{\Delta}_{i,j}\}, \boldsymbol{\Sigma}, \\ \{\mathbf{L}_j\}, \{\lambda_j\}}} \ & -\log |\boldsymbol{\Gamma}| - \log |\boldsymbol{\Sigma}| + \mathrm{tr}\,(\boldsymbol{\Gamma}\boldsymbol{\Sigma}) + \sum_{j=1}^{k} \lambda_j P - \sum_{i=1}^{k} N_i \\ \text{s.t.} \quad & \begin{bmatrix} \lambda_j \mathbf{I} & \mathbf{F}_j^* \\ \mathbf{F}_j & \boldsymbol{\Gamma}^{-1} \end{bmatrix} \succeq 0, \forall j; \boldsymbol{\Sigma} \in \mathcal{T}, \\ & \|\boldsymbol{\Delta}_{i,j}\|_F \leq \epsilon_{i,j}, \forall j, i \leq j. \end{aligned} \tag{6.52}$$

Note that the last constraint in (6.52) can be written as a linear matrix inequality. Next, for any given $\{\mathbf{\Delta}_{i,j}\}$, $\mathbf{\Sigma} \in \mathcal{T}$, $\{\mathbf{L}_j\}$, the optimization problem in the remaining variables can also be shown to be a MAX-DET formulation. Thus, we have an iterative algorithm to compute an outer bound.

6.8 Numerical Results

In this section, we provide few numerical results for the perfect channel scenario. We consider the 3-user GIFC. The noise covariance is normalized to identity matrix. All channel coefficients are chosen as i.i.d. complex Gaussian $\mathcal{CN}(0, 1)$. All plots are averaged over 10^3 channel realizations.

First, consider two antennas at all transmitters and all receivers. In Figure 6.1, first, we plot the sum rate versus per-user power achieved using the iterative algorithm developed for single user decoding. Next, we plot the trivial upper bound obtained by ignoring interference along with a simple lower bound obtained when each user transmits with along the singular-vectors of its direct channel with equal power. As expected, in the latter case the system becomes interference limited at high SNR and the sum-rate saturates. Lastly, we plot the iterative upper bound developed using genie-MACs. In Figure 6.1, for a typical power value like 10 dB, our upper bound is significantly better than the trivial outer bound and "not-far" from our achievable sum rate. We observe in simulations that the iterative achievability algorithms can set $d = \min\{M, N\}$ without reducing performance (in fact this gives the best performance). This is a significant advantage over (explicit) interference alignment schemes that

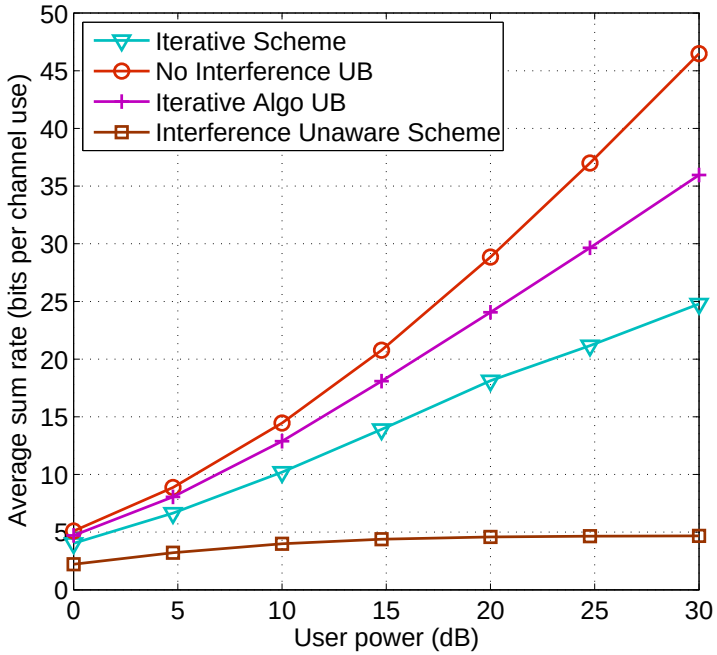

Figure 6.1: Average sum rate for 2×2 3-user GIFC; Comparison with outer bound

need to know "feasibility" (of the assigned DoF vector) a-priori.[6] The gap between the upper bound and the achievable rate seen in numerical results suggest that there is room for making further progress in terms of lower and/or upper bounds. Towards this end, our achievability objective can potentially be appended with objectives for "favoring alignment". Further, the outer bounds can be appended with tighter two-user bounds.

Next, consider four antennas at all transmitters and all receivers. In

[6]We note that with perfect CSI, for any given set of precoders, the rate achieved via single user decoding can also be achieved with single stream decoding after orthogonalizing the "effective" MIMO channels.

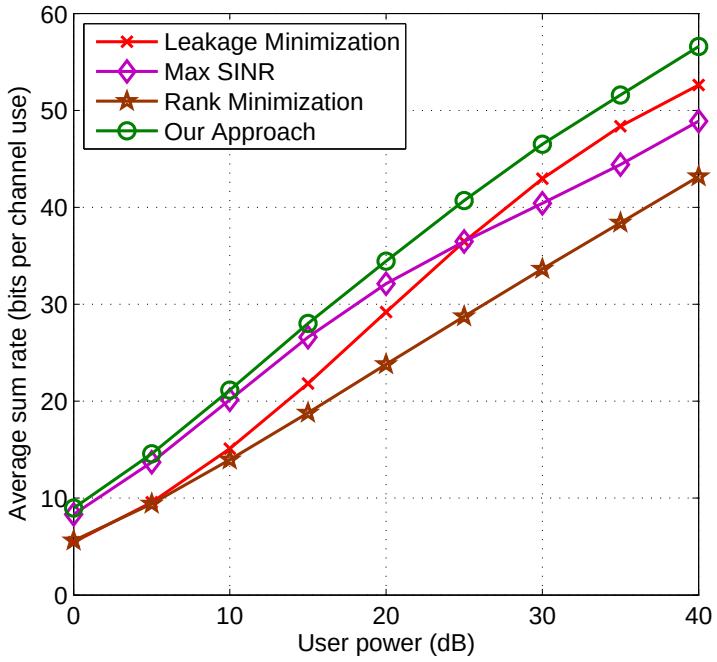

Figure 6.2: Average sum rate for 4×4 3-user GIFC; Comparison with existing schemes

Figure 6.2, we compare our approach with existing schemes. 'Max-SINR' and 'Leakage Minimization' are schemes developed in [45]. 'Rank Minimization' refers to the scheme developed in [97]. In this simulation scenario, our approach matches or outperforms the sum-rate given by existing schemes at all SNRs. Note that the main distinguishing feature of our approach compared to these existing schemes is its applicability to any weighted-sum rate objective even in the presence of channel uncertainty.

Chapter 7

Pilot Contamination and Precoding in Multi-Cell TDD Systems

7.1 Introduction

Multiple antennas, especially at the base-station, have now become an accepted (and in fact, a central) feature of cellular networks. These networks have been studied extensively over the past two decades (see [12] and references therein). It is now well understood that channel state information (CSI) at the base station is an essential component when trying to maximize network throughput. Systems with varying degrees of CSI have been studied in great detail in literature. The primary framework under which these have been studied is frequency division duplex (FDD) systems, where the CSI is typically obtained through (limited) feedback. In Chapter 6, we consider this framework for the K-user MIMO interference channel setting. There is a rich body of work in jointly designing the feedback mechanism with (pre)coding strategies to maximize throughput in MIMO downlink (or broadcast channel) [8, 35, 54, 61, 114, 136]. Time division duplex (TDD) systems, however, have a fundamentally different architecture from the ones studied in FDD systems [65, 66, 84]. Our goal is to develop a clear understanding of mechanisms for acquiring CSI and subsequently designing precoding strategies for

multi-cell TDD systems.

An important distinguishing feature of TDD systems is the notion of *reciprocity*, where the reverse channel is used as an estimate of the forward channel. While utilizing reciprocity, the differences in the transfer characteristics of the amplifiers and the filters in the two directions must be accounted for. Arguably, this reverse channel estimation one of the best advantages of a TDD architecture, as it eliminates the need for feedback, and uplink training together with the reciprocity of the wireless medium [85, 126] is sufficient to provide us with the desired CSI. In [49], channel reciprocity has been validated through experiments. However, as we see next, this channel estimate is not without issues that must be addressed before it proves useful.

7.1.1 Our Approach & Contributions

We consider uplink training and transmit precoding in a multi-cell scenario with L cells, where each cell consists of a base station with M antennas and K users with single antenna each. The impact of uplink training on the resulting channel estimate (and thus system performance) in the multi-cell scenario is significantly different from that in a single-cell scenario. In the multi-cell scenario, non-orthogonal training sequences (pilots) must be utilized, as orthogonal pilots would need to be least $K \times L$ symbols long which is infeasible for large L. Specifically, short channel coherence times due to mobility do not allow for such long training sequences.

This non-orthogonal nature causes *pilot contamination*, which is en-

countered only when analyzing a multi-cell MIMO system with training, and is lost when narrowing focus to a single-cell setting or to a multi-cell setting where channel information is assumed available at no cost. We perform a detailed study of this problem and consider precoding in its presence. Pilot contamination occurs when the channels to users from other cells pollute the channel estimate at the base station in one cell. Thus, our goals are, first, to study the impact of pilot contamination (and thus achievable rates), and then, to develop methods that mitigate this contamination. In older generation cellular systems, multiple factors including large reuse distance of any training signal and randomization in the selection of pilots would have helped in keeping the impact of pilot contamination reasonable. However, with newer generation systems designed for more aggressive reuse of spectrum, this impact is very crucial to understand and mitigate. We note that pilot contamination must also figure in Cooperative MIMO (also called Network MIMO [43, 123]) where clusters of base stations are wired together to create distributed arrays, and where pilots must be re-used over multiple clusters. The fact that pilot contamination hasn't surfaced in FDD studies is primarily due to the assumption of partial CSI with independently corrupted noise as in Chapter 6, and are not fully incorporating the impact of channel estimation.

The fundamental problem associated with pilot contamination is evident even in the simple multi-cell scenario shown in Figure 7.1. Consider two cells $i \in \{1, 2\}$, each consisting of one base station and one user. Let $\mathbf{h}_{ij}$ denote the channel between the base station in the i-th cell and the user

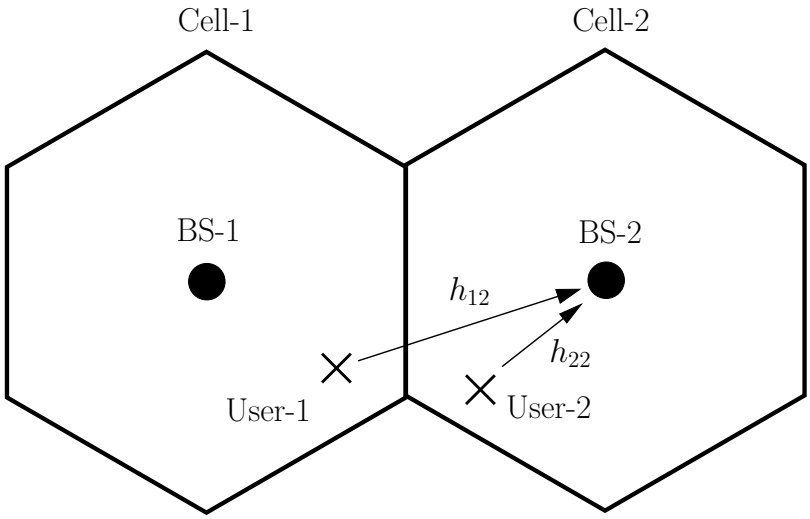

Figure 7.1: A two-cell example with one user in each cell; Both users transmit non-orthogonal pilots during uplink training, which leads to pilot contamination at both the base stations.

in the j-th cell. Let the training sequences used by both the users be same. In this case, the MMSE channel estimate of $\mathbf{h}_{22}$ at the base station in the 2-nd cell is $\hat{\mathbf{h}}_{22} = c_1\mathbf{h}_{12} + c_2\mathbf{h}_{22} + c\mathbf{w}$. Here c_1, c_2 and c are constants that depend on the propagation factors and the transmit powers of mobiles, and $\mathbf{w}$ is $\mathcal{CN}(0, \mathbf{I})$ additive noise. The base station in the 2-nd cell uses this channel estimate to form a precoding vector $\mathbf{a}_2 = f(\hat{\mathbf{h}}_{22})$, which is usually aligned with the channel estimate, that is $\mathbf{a}_2 = \text{const} \cdot \hat{\mathbf{h}}_{22}$. However, by doing this, the base station (partially) aligns the transmitted signal with both $\mathbf{h}_{22}$ (which is desirable) and $\mathbf{h}_{12}$ (which is undesirable). Both signal ($\mathbf{h}_{22}\mathbf{a}_2^\dagger$) and interference ($\mathbf{h}_{12}\mathbf{a}_2^\dagger$) statistically behave similarly. Therefore, the general assumption that the precoding vector used by a base station in one cell is uncorrelated with the channel to users in other cells is not valid with uplink training using

non-orthogonal training sequences.

To perform this analysis, we first develop analytical expressions using techniques similar to those used in [66, 84]. For the setting with one user in every cell, we derive closed-form expressions for achievable rates. These closed-form expressions allow us to determine the extent to which pilot contamination impacts system performance. In particular, we show that the achievable rates can saturate with the number of antennas at the base station M. This analysis will allow system designers to determine the appropriate frequency/time/pilot reuse factor to maximize system throughput in the presence of pilot contamination.

In the multi-cell scenario, there has been significant work on utilizing coordination among base stations [43,111,123,139] when CSI is available. This existing body of work focuses on the gain that can be obtained through coordination of the base stations. Dirty paper coding based approaches and joint beamforming/precoding approaches are considered in [139]. Linear precoding methods for clustered networks with full intra-cluster coordination and limited inter-cluster coordination are proposed in [140]. These approaches generally require "good" channel estimates at the base stations. Due to non-orthogonal training sequences, the resulting channel estimate (of the channel between a base station and all users) can be shown to be rank deficient. We develop a multi-cell MMSE-based precoding method that depends on the set of training sequences assigned to the users. Note that this MMSE-based precoding is for the general setting with multiple users in every cell. Our approach does

not need coordination between base stations required by the joint precoding techniques.[1] When coordination is present, this approach can be applied at the inter-cluster level. The MMSE-based precoding derived has several advantages. In addition to being a linear precoding method, it has a simple closed-form expression that results from an intuitive optimization problem formulation. For many training sequence allocations, numerical results show that our approach gives significant gains over certain popular single-cell precoding methods including zero-forcing precoding.

7.1.2 Related Work

Over the past decade, a variety of aspects of downlink and uplink transmission problems in a single cell setting have been studied. In information theoretic literature, these problems are studied as the broadcast channel (BC) and the multiple access channel (MAC) respectively. For Gaussian BC and general MAC, the problems have been studied for both single and multiple antenna cases. The sum capacity of the multi-antenna Gaussian BC has been shown to be achieved by dirty paper coding (DPC) in [21, 124, 125, 137]. It was shown in [129] that DPC characterizes the full capacity region of the multi-antenna Gaussian BC. These results assume perfect CSI at the base station and the users. In addition, the DPC technique is computationally challenging to implement in practice. There has been significant research focus on reducing the computational complexity at the base station and the users. In this regard,

[1] There is no exchange of channel state information among base stations.

different precoding schemes with low complexity have been proposed. This body of work [5,13,53,115,116] demonstratess that sum rates close to sum capacity can be achieved with much lower computational complexity. However, these results assume perfect CSI at the base station and the users.

The problem of lack of channel CSI is usually studied by considering one of the following two settings. As discussed before, in the first setting, CSI at users is assumed to be available and a limited feedback link is assumed to exist from the users to the base station. In [8, 54, 61, 62, 114, 136] such a setting is considered. In [61], the authors show that at high signal to noise ratios (SNRs), the feedback rate required per user must grow linearly with the SNR (in dB) in order to obtain the full MIMO BC multiplexing gain. The main result in [136] is that the extent of CSI feedback can be reduced by exploiting multi-user diversity. In [8] it is shown that nonrandom vector quantizers can significantly increase the MIMO downlink throughput. In [54], the authors design a joint CSI quantization, beamforming and scheduling algorithm to attain optimal throughput scaling. In the next setting, time-division duplex systems are considered and channel training and estimation error are accounted for in the net achievable rate. This approach is used in [20, 46, 65, 84]. In [84], the authors give a lower bound on sum capacity and demonstrate that it is always beneficial to increase the number of antennas at the base station. In [65], the authors study a heterogeneous user setting and present scheduling and precoding methods for this setting. In [46], the authors consider two-way training and propose two variants of linear MMSE precoders as alternatives to

linear zero-forcing precoder used in [84]. Single-cell analysis of TDD systems are also provided in [20].

Given this extensive body of literature in single-cell systems, our main contribution is in understanding multi-cell systems with channel training. Its emphasis is on TDD systems, which are arguably poorly studied compared to FDD systems. Specifically, the main contributions are to demonstrate the pilot contamination problem associated with uplink training, understand its impact on the operation of multi-cell MIMO TDD cellular systems, and develop a new precoding method to mitigate this problem.

7.1.3 Organization

In Section 7.2, we describe the multi-cell system model. In Section 7.3, we explain the communication scheme and the technique to obtain achievable rates. We analyze the effect of pilot contamination in Section 7.4, and give the details of the new precoding method in Section 7.5. We present few numerical results in Section 7.6. The proofs of the theorems are given in Appendix C.

7.2 Multi-Cell TDD System Model

We consider a cellular system with L cells numbered $1, 2, \cdots, L$. Each cell consists of one base station with M antennas and $K(\leq M)$ single-antenna users. Let the average power (during transmission) at the base station be p_f and the average power (during transmission) at each user be p_r. The propagation factor between the m-th base station antenna of the l-th cell

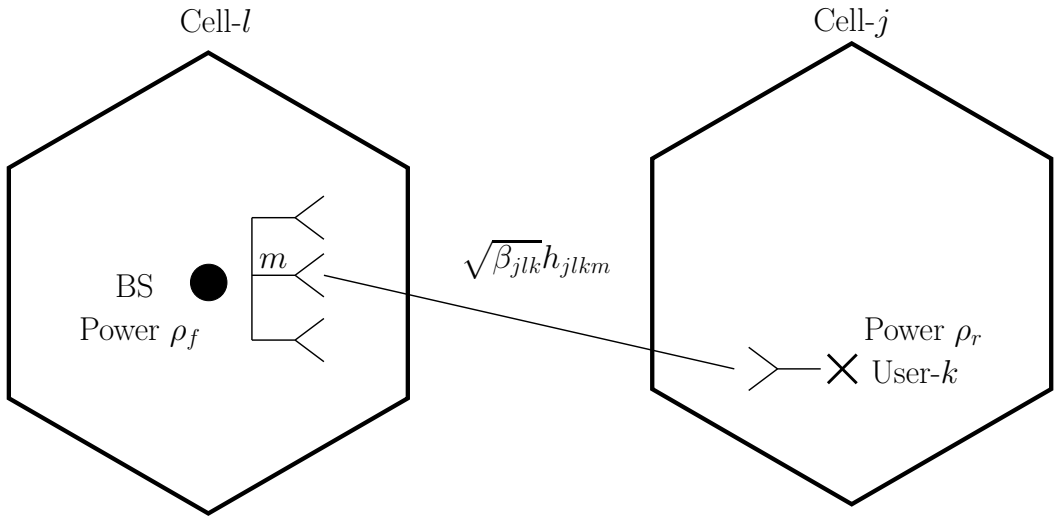

Figure 7.2: System model showing the base station in l-th cell and the k-th user in j-th cell

and the k-th user of the j-th cell is $\sqrt{\beta_{jlk}}h_{jlkm}$,[2] where $\{\beta_{jlk}\}$ are non-negative constants and assumed to be known to everybody, and $\{h_{jlkm}\}$ are independent and identically distributed (i.i.d.) zero-mean, circularly-symmetric complex Gaussian $\mathcal{CN}(0,1)$ random variables and known to nobody. This system model is shown in Figure 7.2. The above assumptions are fairly accurate and justified due to the following reason. The $\{\beta_{jlk}\}$ values model path-loss and shadowing that change slowly and can be learned over long period of time, while the $\{h_{jlkm}\}$ values model fading that change relatively fast and must be learned and used very quickly. Since the cell layout and shadowing are captured using the constant $\{\beta_{jlk}\}$ values, for the purpose of analysis, the specific details of the

[2]For compact notation, we do not separate the subscript or superscript indices using commas.

cell layout and shadowing model are irrelevant. In other words, any cell layout and any shadowing model can be incorporated with the above abstraction.

We assume channel reciprocity for the forward and reverse links, i.e., the propagation factor $\sqrt{\beta_{jlk}}h_{jlkm}$ is same for both forward and reverse links, and block fading, i.e., $\{h_{jlkm}\}$ remains constant for a duration of T symbols. Note that we allow for a constant factor variation in forward and reverse propagation factors through the different average power constraints at the base stations and the users. The additive noises at all terminals are i.i.d. $\mathcal{CN}(0,1)$ random variables. The system equations describing the signals received at the base station and the users are given in the next section.

7.3 Communication Scheme

The communication scheme consists of two phases: uplink training and data transmission. Uplink training phase consists of users transmitting training pilots, and base stations obtaining channel estimates. Data transmission phase consists of base stations transmitting data to the users through transmit precoding. Next, we describe these phases briefly and provide a set of achievable data rates using a given precoding method.

7.3.1 Uplink Training

At the beginning of every coherence interval, all users (in all cells) transmit training sequences, which are τ length column vectors.[3] Let $\sqrt{\tau}\boldsymbol{\psi}_{jk}$ (normalized such that $\boldsymbol{\psi}_{jk}^{\dagger}\boldsymbol{\psi}_{jk} = 1$) be the training vector transmitted by the k-th user in the j-th cell. Consider the base station of the l-th cell. The τ length column vector received at the m-th antenna of this base station is

$$\mathbf{y}_{lm} = \sum_{j=1}^{L}\sum_{k=1}^{K}\sqrt{p_r\tau\beta_{jlk}}h_{jlkm}\boldsymbol{\psi}_{jk} + \mathbf{w}_{lm}, \tag{7.1}$$

where $\mathbf{w}_{lm}$ is the additive noise. Let $\mathbf{Y}_l = [\mathbf{y}_{l1}\ \mathbf{y}_{l2}\ \cdots\ \mathbf{y}_{lM}]$ ($\tau \times M$ matrix), $\mathbf{W}_l = [\mathbf{w}_{l1}\ \mathbf{w}_{l2}\ \cdots\ \mathbf{w}_{lM}]$ ($\tau \times M$ matrix), $\boldsymbol{\Psi}_j = [\boldsymbol{\psi}_{j1}\ \boldsymbol{\psi}_{j2}\ \cdots\ \boldsymbol{\psi}_{jK}]$ ($\tau \times K$ matrix), $\mathbf{D}_{jl} = \mathrm{diag}\{[\beta_{jl1}\ \beta_{jl2}\ \cdots\ \beta_{jlK}]\}$, and

$$\mathbf{H}_{jl} = \begin{bmatrix} h_{jl11} & \cdots & h_{jl1M} \\ \vdots & \ddots & \vdots \\ h_{jlK1} & \cdots & h_{jlKM} \end{bmatrix}.$$

From (7.1), the signal received at this base station can be expressed as

$$\mathbf{Y}_l = \sqrt{p_r\tau}\sum_{j=1}^{L}\left(\boldsymbol{\Psi}_j\mathbf{D}_{jl}^{\frac{1}{2}}\mathbf{H}_{jl}\right) + \mathbf{W}_l. \tag{7.2}$$

The MMSE estimate of the channel $\mathbf{H}_{il}$ given $\mathbf{Y}_l$ in (7.2) is

$$\hat{\mathbf{H}}_{jl} = \sqrt{p_r\tau}\mathbf{D}_{jl}^{\frac{1}{2}}\boldsymbol{\Psi}_j^{\dagger}\left(\mathbf{I} + p_r\tau\sum_{i=1}^{L}\boldsymbol{\Psi}_i\mathbf{D}_{il}\boldsymbol{\Psi}_i^{\dagger}\right)^{-1}\mathbf{Y}_l. \tag{7.3}$$

This MMSE estimate in (7.3) follows from standard results in estimation theory (for example see [72]). We denote the MMSE estimate of the channel

[3] We assume that there is time synchronization present in the system for coherent uplink transmission.

between this base station and all users by $\hat{\mathbf{H}}_l = [\hat{\mathbf{H}}_{1l}\, \hat{\mathbf{H}}_{2l}\, \cdots\, \hat{\mathbf{H}}_{Ll}]$. This notation is used later in Section 7.4.

7.3.2 Downlink Transmission

Consider the base station of the l-th cell. Let the information symbols to be transmitted to users in the l-th cell be $\mathbf{q}_l = [q_{l1}\ q_{l2}\ \cdots\ q_{lK}]^T$ and the $M \times K$ linear precoding matrix be $\mathbf{A}_l = f(\hat{\mathbf{H}}_l)$. The function $f(\cdot)$ corresponds to the specific (linear) precoding method performed at the base station. The signal vector transmitted by this base station is $\mathbf{A}_l\mathbf{q}_l$. We consider transmission symbols and precoding methods such that $\mathbb{E}[\mathbf{q}_l] = \mathbf{0}$, $\mathbb{E}[\mathbf{q}_l\mathbf{q}_l^\dagger] = \mathbf{I}$, and $\mathrm{tr}(\mathbf{A}_l^\dagger\mathbf{A}_l) = 1$. These (sufficient) conditions imply that the average power constraint at the base station is satisfied.

Now, consider the users in the j-th cell. The noisy signal vector received by these users is

$$\mathbf{x}_j = \sum_{l=1}^{L} \sqrt{p_f}\mathbf{D}_{jl}^{\frac{1}{2}}\mathbf{H}_{jl}\mathbf{A}_l\mathbf{q}_l + \mathbf{z}_j, \tag{7.4}$$

where $\mathbf{z}_j$ is the additive noise. From (7.4), the signal received by the k-th user can be expressed as

$$x_{jk} = \sum_{l=1}^{L}\sum_{i=1}^{K} \sqrt{p_f\beta_{jlk}}[h_{jlk1}\ h_{jlk2}\ \cdots\ h_{jlkM}]\mathbf{a}_{li}q_{li} + z_{jk}, \tag{7.5}$$

where $\mathbf{a}_{li}$ is the i-th column of the precoding matrix $\mathbf{A}_l$ and z_{jk} is the k-th element of $\mathbf{z}_j$.

7.3.3 Achievable Rates

Next, we provide a set of achievable rates using the method suggested in [84]. With the above communication scheme, the base stations have channel estimates while the users do not have any channel estimate. Therefore, the achievable rates we derive have a different structure compared to typical rate expressions. In particular, the effective noise term has channel variations around the mean in addition to typical terms.

Let $g_{li}^{jk} = \sqrt{p_f \beta_{jlk}}\,[h_{jlk1} \cdots h_{jlkM}]\,\mathbf{a}_{li}$. Now, (7.5) can be written in the form

$$\begin{aligned} x_{jk} &= \sum_{l=1}^{M}\sum_{i=1}^{K} g_{li}^{jk} q_{li} + z_{jk}, \\ &= \mathbb{E}\left[g_{jk}^{jk}\right] q_{jk} + \left(g_{jk}^{jk} - \mathbb{E}\left[g_{jk}^{jk}\right]\right) q_{jk} + \sum_{(l,i)\neq(j,k)} g_{li}^{jk} q_{li} + z_{jk}. \end{aligned} \quad (7.6)$$

In (7.6), the effective noise is defined as

$$z_{jk}^{'} = \left(g_{jk}^{jk} - \mathbb{E}\left[g_{jk}^{jk}\right]\right) q_{jk} + \sum_{(l,i)\neq(j,k)} g_{li}^{jk} q_{li} + z_{jk}. \quad (7.7)$$

Now, the effective point-to-point channel described by (7.6) can be written in the familiar form

$$x_{jk} = \mathbb{E}\left[g_{jk}^{jk}\right] q_{jk} + z_{jk}^{'}, \quad (7.8)$$

where q_{jk} is the input, x_{jk} is the output, $\mathbb{E}\left[g_{jk}^{jk}\right]$ is the known channel and $z_{jk}^{'}$ is the additive noise. $\mathbb{E}\left[g_{jk}^{jk}\right]$ is known as it only depends on the channel distribution and not the instantaneous channel. However, the additive noise

is neither independent nor Gaussian. We use the result in [52] that shows that worst-case uncorrelated additive noise is independent Gaussian noise of same variance to derive the following achievable rates.

Theorem 7.1. *Consider the point-to-point communication channels given by (7.8). Then, the following set of rates are achievable:*

$$R_{jk} = C\left(\frac{\left|\mathbb{E}\left[g_{jk}^{jk}\right]\right|^2}{1+\text{var}\left\{g_{jk}^{jk}\right\}+\sum_{(l,i)\neq(j,k)}\mathbb{E}\left[\left|g_{li}^{jk}\right|^2\right]}\right), \tag{7.9}$$

where $C(\theta) = \log_2(1+\theta)$.

Proof. Please see Appendix C.1. □

Remark 7.1. *The set of achievable rates given by (7.9) is valid for any linear precoding method, and depends on the precoding method through the expectation and variance terms appearing in (7.9).*

Similar achievable rates are used in the single-cell setting as well to study and/or compare precoding methods. Next, we perform pilot contamination analysis using these achievable rates.

7.4 Pilot Contamination Analysis

We analyze the pilot contamination problem in the following setting: one user per cell ($K = 1$), same training sequence used by all users ($\boldsymbol{\psi}_{j1} = \boldsymbol{\psi}, \forall j$) and matched-filter (MF) precoding. We consider this setting as it captures the primary effect of pilot contamination which is the correlation between

the precoding matrix (vector in this setting) used by the base station in a cell and channel to users in other cells. We provide simple and insightful analytical results in this setting. As mentioned earlier, we emphasize that the pilot contamination problem results from uplink training with non-orthogonal training sequences, and hence, it is not specific to the setting considered here. However, the level of its impact on the achievable rates would vary depending on the system settings.

In order to simplify notation, we drop the subscripts associated with the users in every cell. In this section, $\mathbf{H}_{jl}$, $\hat{\mathbf{H}}_{jl}$ and $\mathbf{A}_l$ are vectors and we denote these using $\mathbf{h}_{jl}$, $\hat{\mathbf{h}}_{jl}$ and $\mathbf{a}_l$, respectively. The matched-filter used at the base station in the l-th cell is given by $\mathbf{a}_l = \hat{\mathbf{h}}_{ll}^{\dagger}/\|\hat{\mathbf{h}}_{ll}\|$. The user in the j-th cell receives signal from its base station and from other base stations. From (7.4), this received signal is

$$x_{jk} = \sqrt{p_f \beta_{jj}}\mathbf{h}_{jj}\mathbf{a}_j q_j + \sum_{l \neq j} \sqrt{p_f \beta_{jl}}\mathbf{h}_{jl}\mathbf{a}_l q_l + z_j. \tag{7.10}$$

We compute first and second order moments of the effective channel gain and the inter-cell interference and use these to obtain a simple expression for the achievable rate given by (7.9).

In the setting considered here, the MMSE estimate of $\mathbf{h}_{jl}$ based on $\mathbf{Y}_l$ given by (7.3) can be simplified using matrix inversion lemma and the fact

that $\boldsymbol{\psi}^\dagger\boldsymbol{\psi} = 1$ as follows:

$$\begin{aligned}
\hat{\mathbf{h}}_{jl} &= \sqrt{p_r\tau\beta_{jl}}\boldsymbol{\psi}^\dagger\left(\mathbf{I} + \boldsymbol{\psi}\left(p_r\tau\sum_{i=1}^{L}\beta_{il}\right)\boldsymbol{\psi}^\dagger\right)^{-1}\mathbf{Y}_l, \\
&= \sqrt{p_r\tau\beta_{jl}}\boldsymbol{\psi}^\dagger\left(\mathbf{I} - \frac{\boldsymbol{\psi}\left(p_r\tau\sum_{i=1}^{L}\beta_{il}\right)\boldsymbol{\psi}^\dagger}{1+p_r\tau\sum_{i=1}^{L}\beta_{il}}\right)\mathbf{Y}_l, \\
&= \sqrt{p_r\tau\beta_{jl}}\left(\boldsymbol{\psi}^\dagger - \frac{\left(p_r\tau\sum_{i=1}^{L}\beta_{il}\right)}{1+p_r\tau\sum_{i=1}^{L}\beta_{il}}\boldsymbol{\psi}^\dagger\right)\mathbf{Y}_l, \\
&= \frac{\sqrt{p_r\tau\beta_{jl}}}{1+p_r\tau\sum_{i=1}^{L}\beta_{il}}\boldsymbol{\psi}^\dagger\mathbf{Y}_l.
\end{aligned}$$

Remark 7.2. *The above channel estimates clearly suggest the graveness of the pilot contamination impairment. For a given base station, its estimate of every channel is simply a scaled version of the same vector $\boldsymbol{\psi}^\dagger\mathbf{Y}_l$. Thus, it cannot distinguish between the channel to its user and other users, which makes pilot contamination a fundamental problem in multi-cell systems.*

Since $\boldsymbol{\psi}^\dagger\mathbf{Y}_l$ is proportional to the MMSE estimate of $\mathbf{h}_{jl}$ for any j, we have

$$\frac{\hat{\mathbf{h}}_{jl}^\dagger}{\|\hat{\mathbf{h}}_{jl}\|} = \frac{\mathbf{Y}_l^\dagger\boldsymbol{\psi}}{\|\boldsymbol{\psi}^\dagger\mathbf{Y}_l\|}, \forall j. \tag{7.11}$$

Using (7.11), we obtain

$$\begin{aligned}
\mathbf{h}_{jl}\mathbf{a}_l &= \mathbf{h}_{jl}\frac{\hat{\mathbf{h}}_{jl}^\dagger}{\left\|\hat{\mathbf{h}}_{jl}^\dagger\right\|}, \\
&= \left\|\hat{\mathbf{h}}_{jl}^\dagger\right\| + \tilde{\mathbf{h}}_{jl}\frac{\hat{\mathbf{h}}_{jl}^\dagger}{\left\|\hat{\mathbf{h}}_{jl}^\dagger\right\|},
\end{aligned} \tag{7.12}$$

where $\tilde{\mathbf{h}}_{jl} = \mathbf{h}_{jl} - \hat{\mathbf{h}}_{jl}$. From the properties of MMSE estimation, we know that $\hat{\mathbf{h}}_{jl}$ is independent of $\tilde{\mathbf{h}}_{jl}$,

$$\hat{\mathbf{h}}_{jl} \sim \mathcal{CN}\left(\mathbf{0}, \frac{p_r \tau \beta_{jl}}{1 + p_r \tau \sum_{i=1}^{L} \beta_{il}} \mathbf{I}\right),$$

and

$$\tilde{\mathbf{h}}_{jl} \sim \mathcal{CN}\left(\mathbf{0}, \frac{1 + p_r \tau \sum_{i \neq j} \beta_{il}}{1 + p_r \tau \sum_{i=1}^{L} \beta_{il}} \mathbf{I}\right).$$

These results are used next.

From (7.12), we get

$$\begin{aligned} \mathbb{E}\left[\mathbf{h}_{jl} \mathbf{a}_l\right] &= \mathbb{E}\left[\left\|\hat{\mathbf{h}}_{jl}^{\dagger}\right\|\right], \\ &= \sqrt{\frac{p_r \tau \beta_{jl}}{1 + p_r \tau \sum_{i=1}^{L} \beta_{il}}} \mathbb{E}\left[\theta\right], \end{aligned} \tag{7.13}$$

where $\theta = \sqrt{\sum_{m=1}^{M} |u_m|^2}$ and $\{u_m\}$ is i.i.d. $\mathcal{CN}(0,1)$. From (7.12), we also have

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{h}_{jl} \mathbf{a}_l\|^2\right] &= \mathbb{E}\left[\left\|\hat{\mathbf{h}}_{jl}^{\dagger}\right\|^2\right] + \mathbb{E}\left[\frac{\hat{\mathbf{h}}_{jl}}{\left\|\hat{\mathbf{h}}_{jl}^{\dagger}\right\|} \tilde{\mathbf{h}}_{jl}^{\dagger} \tilde{\mathbf{h}}_{jl} \frac{\hat{\mathbf{h}}_{jl}^{\dagger}}{\left\|\hat{\mathbf{h}}_{jl}^{\dagger}\right\|}\right], \\ &= \frac{p_r \tau \beta_{jl}}{1 + p_r \tau \sum_{i=1}^{L} \beta_{il}} \mathbb{E}[\theta^2] + \frac{1 + p_r \tau \sum_{i \neq j} \beta_{il}}{1 + p_r \tau \sum_{i=1}^{L} \beta_{il}}. \end{aligned} \tag{7.14}$$

Next, we state two lemmas required to obtain a closed-form expression for the achievable rate.

Lemma 7.2. *The effective channel gain in (7.10) has expectation*

$$\mathbb{E}\left[\sqrt{p_f \beta_{jj}} \mathbf{h}_{jj} \mathbf{a}_j\right] = \left(p_f \beta_{jj} \frac{p_r \tau \beta_{jj}}{1 + p_r \tau \sum_{i=1}^{L} \beta_{ij}}\right)^{1/2} \mathbb{E}[\theta]$$

and variance

$$\mathrm{var}\left\{\sqrt{p_f\beta_{jj}}\mathbf{h}_{jj}\mathbf{a}_j\right\} = p_f\beta_{jj}\left(\frac{p_r\tau\beta_{jj}}{1+p_r\tau\sum_{i=1}^{L}\beta_{ij}}\,\mathrm{var}\{\theta\} + \frac{1+p_r\tau\sum_{i\neq j}\beta_{ij}}{1+p_r\tau\sum_{i=1}^{L}\beta_{ij}}\right).$$

Proof. The proof follows from (7.13) and (7.14). Note that $\mathrm{var}\{\theta\} = \mathbb{E}[\theta^2] - (\mathbb{E}[\theta])^2$ by definition. □

Lemma 7.3. *For both signal and interference terms in (7.10), the first and second order moments are as follows:*

$$\mathbb{E}\left[\sqrt{p_f\beta_{jl}}\mathbf{h}_{jl}\mathbf{a}_l q_l\right] = 0,$$

$$\mathbb{E}\left[\left|\sqrt{p_f\beta_{jl}}\mathbf{h}_{jl}\mathbf{a}_l q_l\right|^2\right] = p_f\beta_{jl}\left(\frac{p_r\tau\beta_{jl}}{1+p_r\tau\sum_{i=1}^{L}\beta_{il}}\mathbb{E}[\theta^2] + \frac{1+p_r\tau\sum_{i\neq j}\beta_{il}}{1+p_r\tau\sum_{i=1}^{L}\beta_{il}}\right).$$

Proof. Since $\mathbb{E}[q_l] = 0$ and q_l is independent of $\mathbf{h}_{jl}$ and $\mathbf{a}_l$, it is clear that

$$\mathbb{E}\left[\sqrt{p_f\beta_{jl}}\mathbf{h}_{jl}\mathbf{a}_l q_l\right] = 0.$$

The proof of the second order moment follows directly from (7.14). □

The main result of this section is given in the next theorem. This theorem provides a closed-form expression for the achievable rates under the setting considered in this section, i.e., one user per cell ($K = 1$), same training sequence used by all users ($\boldsymbol{\psi}_{j1} = \boldsymbol{\psi}, \forall j$) and matched-filter (MF) precoding.

Theorem 7.4. *For the setting considered, the achievable rate of the user in the j-th cell R_j during downlink transmission in (7.9) is given by*

$$C\left(\frac{p_f\beta_{jj}\frac{p_r\tau\beta_{jj}}{\kappa_j}\mathbb{E}^2[\theta]}{\sum_{l\neq j}p_f\beta_{jl}\frac{p_r\tau\beta_{jl}}{\kappa_l}\mathbb{E}[\theta^2]+\zeta}\right), \tag{7.15}$$

where

$$\zeta = 1 + p_f \beta_{jj} \frac{p_r \tau \beta_{jj}}{\kappa_j} \operatorname{var}\{\theta\} + \sum_{l=1}^{L} p_f \beta_{jl} \frac{1 + p_r \tau \sum_{i \neq j} \beta_{il}}{\kappa_l},$$

$\kappa_j = 1 + p_r \tau \sum_{i=1}^{L} \beta_{ij}$, $\mathbb{E}[\theta] = \frac{\Gamma(M+\frac{1}{2})}{\Gamma(M)}$, $\mathbb{E}[\theta^2] = M$ *and* $\operatorname{var}\{\theta\} = M - \mathbb{E}^2[\theta]$. *Here,* $\Gamma(\cdot)$ *is the Gamma function. For large M, the following limiting expression for achievable rate can be obtained:*

$$\lim_{M \to \infty} R_j = C \left(\frac{\frac{\beta_{jj}^2}{1 + p_r \tau \sum_{i=1}^{L} \beta_{ij}}}{\sum_{l \neq j} \frac{\beta_{jl}^2}{1 + p_r \tau \sum_{i=1}^{L} \beta_{il}}} \right). \tag{7.16}$$

Proof. Please see Appendix C.2. □

For large M, the value of $\operatorname{var}\{\theta\}$ ($\approx 1/4$) is insignificant compared to M. The results of the above theorem show that the performance does saturate with M. Typically, the reverse link is interference-limited, i.e., $p_r \tau \sum_{i=1}^{L} \beta_{il} \gg 1, \forall j$. The term $\sum_{i=1}^{L} \beta_{il}$ is the expected sum of squares of the propagation coefficients between the base station in the j-th cell and all users. Therefore, $\sum_{i=1}^{L} \beta_{il}$ is generally constant with respect to j. Using these approximations in (7.16), we get

$$R_j \approx C \left(\frac{\beta_{jj}^2}{\sum_{l \neq j} \beta_{jl}^2} \right).$$

This clearly show that the impact of pilot contamination can be very significant if cross gains (between cells) are of the same order of direct gains (within the same cell). It suggests frequency/time reuse and pilot reuse techniques to reduce the cross gains (in the same frequency/time) relative to the direct gains.

Remark 7.3. *Our result in Theorem 7.4 is not an asymptotic result. The expression in (7.15) is exact for any value of the number of antennas M at the base stations. Hence, this expression can be used to find the appropriate frequency/time reuse scheme for any given value of M and other system parameters. We do not focus on this, as this would depend largely on the actual system parameters including the cell layout and the shadowing model.*

Remark 7.4. *The result in Theorem 7.4 is for the setting with one user per cell. In the general setting with K users per cell, a similar analysis can be performed, however, it need not simplify to a simple closed-form expression. The achievable rate in (7.9) can be numerically evaluated in the general setting, and this can be used to numerically study the impact of pilot contamination.*

To summarize, the impact of uplink training with non-orthogonal pilots can be serious when the cross-gains are not small compared to the direct gains. This pilot contamination problem is often neglected in theory and even in many large-scale simulations. The analysis in this section shows the need to account for this impact especially in systems with high reuse of training sequences. In addition to uplink training in TDD systems, the pilot contamination problem would appear in other scenarios as well as it is fundamental to training with non-orthogonal pilots.

Next, we proceed to develop a new precoding method referred to as the multi-cell MMSE-based precoding.

7.5 Multi-Cell MMSE-Based Precoding

In the previous section, we show that pilot contamination severely impacts the system performance by increasing the inter-cell interference. In particular, we show that the inter-cell interference grows like the intended signal with the number of antennas M at the base stations while using zero-forcing precoding. Therefore, in the presence of pilot contamination, in addition to frequency/time/pilot reuse schemes, it is crucial to account for inter-cell interference while designing a precoding method. Furthermore, since pilot contamination is originating from the non-orthogonal training sequences, it is important to account for the training sequence allocation while designing a precoding method. The approach of accounting for inter-cell interference while designing a precoding method is common, while the approach of accounting for the training sequence allocation is not. Again, the usual approach is to decouple the channel estimation and precoding completely. However, while using non-orthogonal pilots, this is not the right approach. These observations follow from our pilot contamination analysis in the previous section.

The precoding problem cannot be directly formulated as a joint optimization problem as different base stations have different received training signals. In other words, the problem is decentralized in nature. Therefore, one approach is to apply single-cell precoding methods. For example, since we assume orthogonal training sequences in every cell, we can perform zero-forcing on the users in every cell. The precoding matrix corresponding to this

zero-forcing approach is given by

$$\mathbf{A}_l = \frac{\hat{\mathbf{G}}_{ll}^{\dagger}\left(\hat{\mathbf{G}}_{ll}\hat{\mathbf{G}}_{ll}^{\dagger}\right)^{-1}}{\sqrt{\mathrm{tr}\left[\left(\hat{\mathbf{G}}_{ll}\hat{\mathbf{G}}_{ll}^{\dagger}\right)^{-1}\right]}}, \tag{7.17}$$

where $\hat{\mathbf{G}}_{ll} = \sqrt{p_f}\mathbf{D}_{ll}^{\frac{1}{2}}\hat{\mathbf{H}}_{ll}$. However, this zero-forcing precoding or other single-cell precoding methods do not account for the training sequence allocation, which is potentially the right approach to mitigate the pilot contamination problem. We explore this next.

In order to determine the precoding matrices, we formulate an optimization problem for each precoding matrix. Consider the j-th cell. The signal received by the users in this cell given by (7.4) is a function of all the precoding matrices (used at all the base stations). Therefore, the MMSE-based precoding methods for single-cell setting considered in [46] does not extend (directly) to this setting. Let us consider the signal and interference terms corresponding to the base station in the l-th cell. Based on these terms, we formulate the following optimization problem to obtain the precoding matrix $\mathbf{A}_l$. We use the following notation: $\mathbf{F}_{jl} = \sqrt{p_f}\mathbf{D}_{jl}^{\frac{1}{2}}\mathbf{H}_{jl}$, $\hat{\mathbf{F}}_{jl} = \sqrt{p_f}\mathbf{D}_{jl}^{\frac{1}{2}}\hat{\mathbf{H}}_{jl}$ and $\tilde{\mathbf{F}}_{jl} = \mathbf{F}_{jl} - \hat{\mathbf{F}}_{jl}$ for all j and l. The optimization problem is:

$$\min_{\mathbf{A}_l,\alpha_l} \mathbb{E}_{\tilde{\mathbf{F}}_{jl},\mathbf{z}_l,\mathbf{q}_l}\left[\left\|\alpha_l(\mathbf{F}_{ll}\mathbf{A}_l\mathbf{q}_l + \mathbf{z}_l) - \mathbf{q}_l\right\|^2 + \sum_{j\neq l}\left\|\alpha_l\gamma(\mathbf{F}_{jl}\mathbf{A}_l\mathbf{q}_l)\right\|^2 \middle| \hat{\mathbf{F}}_{jl}\right] \tag{7.18}$$

subject to $\mathrm{tr}(\mathbf{A}_l^{\dagger}\mathbf{A}_l) = 1$. This objective function is very intuitive. The objective function of the problem (7.18) consists of two parts: (i) the sum of squares of "errors" seen by the users in the l-th cell, and (ii) the sum of squares of

interference seen by the users in all other cells. The parameter γ of the optimization problem "controls" the relative weights associated with these two parts. The real scalar parameter α_l is important as it "virtually" corresponds to the potential scaling that can be performed at the users. The optimal solution to the problem (7.18) denoted by $\mathbf{A}_l^{opt}$ is the multi-cell MMSE-based precoding matrix.

The main result of this section is given by the following theorem. This theorem provides a closed-form expression for the multi-cell MMSE-based precoding matrix.

Theorem 7.5. *The optimal solution to the problem (7.18) is*

$$\mathbf{A}_l^{opt} = \frac{1}{\alpha_l^{opt}} \left(\hat{\mathbf{F}}_{ll}^{\dagger} \hat{\mathbf{F}}_{ll} + \gamma^2 \sum_{j \neq l} \hat{\mathbf{F}}_{jl}^{\dagger} \hat{\mathbf{F}}_{jl} + \eta \mathbf{I}_M \right)^{-1} \hat{\mathbf{F}}_{ll}^{\dagger}, \tag{7.19}$$

where

$$\eta = \delta_{ll} + \gamma^2 \sum_{j \neq l} \delta_{jl} + K,$$

δ_{jl} is given by (C.9) and α_l^{opt} is such that $\mathrm{tr}\left(\left(\mathbf{A}_l^{opt} \right)^{\dagger} \mathbf{A}_l^{opt} \right) = 1$.

Proof. Please see Appendix C.3 □

The precoding described above is primarily suited for maximizing the minimum of the rates achieved by all the users. This is because all users are treated equally without differentiating them based on the channels. Therefore, when the performance metric of interest is sum rate, this precoding can be combined with power control, scheduling, and other similar techniques to

enhance the net performance. Since our main concern is the inter-cell interference resulting from pilot contamination, and to avoid too complicated systems, we do not use this possibility. In the next section, all numerical results and comparisons are performed without power control.

7.6 Numerical Results

Multi-Cell MMSE precoding denotes the new precoding method given in (7.19) with parameter γ set to unity.[4] ZF precoding denotes the popular zero-forcing precoding given in (7.17). GPS denotes the single-cell precoding method suggested in [46], which is a special case of the precoding given in (7.19) with parameter γ set to zero. In all the plots, we average the performance metric over 10^3 i.i.d. channel realizations.

7.6.1 Two-cell System

We consider a basic two-cell example to understand the impact of pilot contamination on the total system throughput (sum rate). In particular, we consider a two-cell system with $p_f = 20$ dB, $p_r = 10$ dB and $K = 4$ users in both cells. We set all the direct gains to 1 and all cross gains to a, i.e., for all k, $\beta_{jlk} = 1$ if $j = l$ and $\beta_{jlk} = a$ if $j \neq l$. We capture the main observations using two plots.

First, in Figure 7.3, we use cross gain value of $a = 0.8$ and plot sum rate

[4] In our simulations, we have observed that the performance is not that sensitive to the value of this parameter γ.

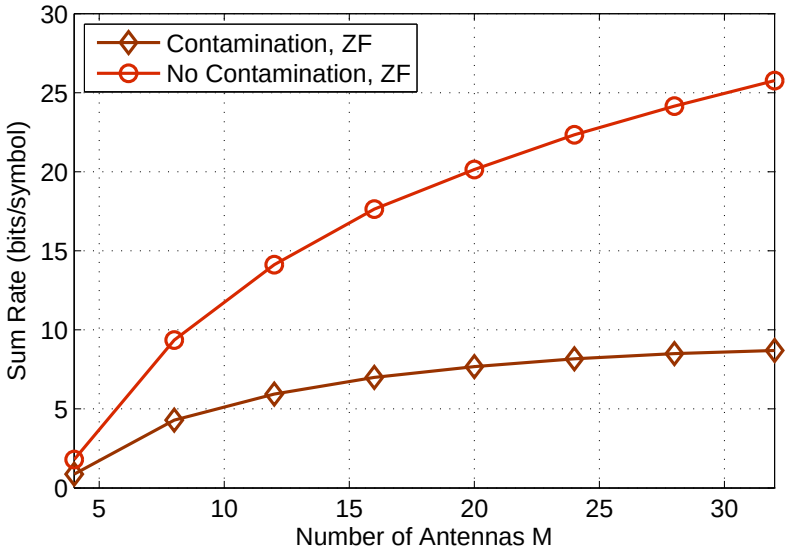

Figure 7.3: Sum rate with and without pilot contamination for the two-cell system; Sum rate saturates in the presence of contamination

versus number of antennas M with zero-forcing for training lengths of $\tau = 4$ (scenario with pilot contamination) and $\tau = 8$ (scenario without contamination). With $\tau = 4$, the orthogonal training sequences used in the 1-st is reused in the 2-nd cell. With $\tau = 8$, all users in the system are given orthogonal pilots. In Figure 7.3, we can clearly observe the saturation of total throughput in the presence of pilot contamination. Note that both scenarios deal with interference. In many practical systems, we cannot necessarily keep τ large ($\tau = 8$ in this example) as the coherence interval is typically very short, which requires using small τ.

In the presence of pilot contamination ($\tau = 4$), neither GPS nor multi-cell MMSE provide noticeable improvement in total throughput (except for

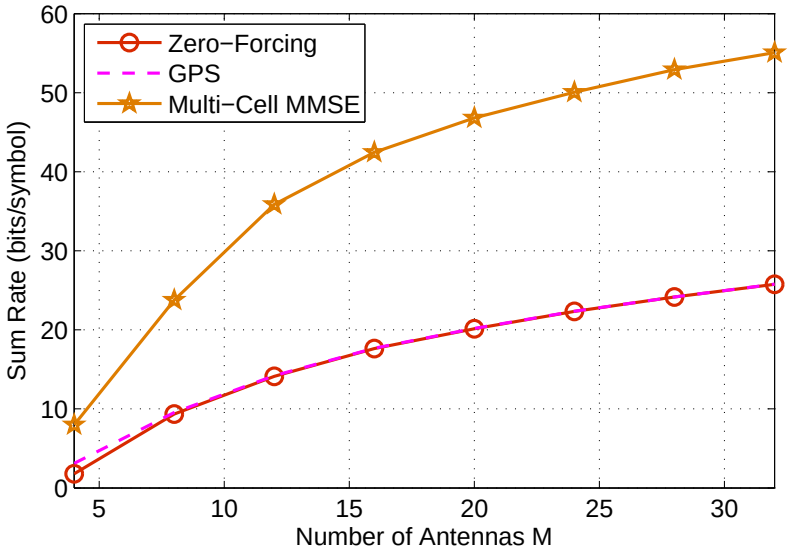

Figure 7.4: Comparison of schemes with orthogonal pilots; Multi-cell MMSE performs inter-cell interference mitigation leading to improved sum rate

small values of M). This is not surprising as pilot contamination is an indispensable[5] problem in this example. For $M = 4$ to $M = 12$, the improvement using GPS and multi-cell MMSE is reasonable. This improvement results from using MMSE instead of zero-forcing, and hence both GPS and multi-cell MMSE provide very similar performance.[6]

In the absence of pilot contamination ($\tau = 8$), the story is different as shown in Figure 7.4. Multi-cell MMSE outperforms both the single-cell schemes (GPS and zero-forcing) by a huge margin. This is possible as multi-cell MMSE is capable of performing efficient inter-cell interference mitigation.

[5]We can overcome pilot contamination using frequency/time reuse.
[6]This plot is not provided as it is not the main focus.

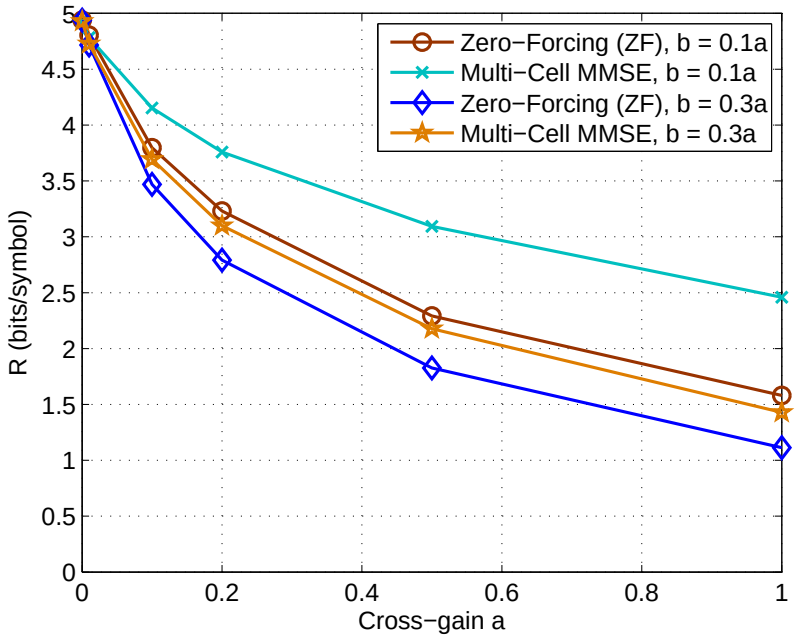

Figure 7.5: Comparison of ZF and multi-cell MMSE; a and b correspond to different cross-gains and R denotes the minimum rate achieved by all users

Note that this inter-cell interference mitigation is achieved using the channel estimates obtained in a distributed manner. This clearly shows the advantage of multi-cell MMSE. However, this example only focuses on the scenario without pilot contamination. Therefore, a natural question is whether multi-cell MMSE can provide throughput gains in a mixed scenario, which is addressed next.

7.6.2 Multi-cell System

We consider a multi-cell system with $L = 4$ cells, $M = 8$ antennas at all base stations, $K = 2$ users in every cell and training length of $\tau = 4$. We

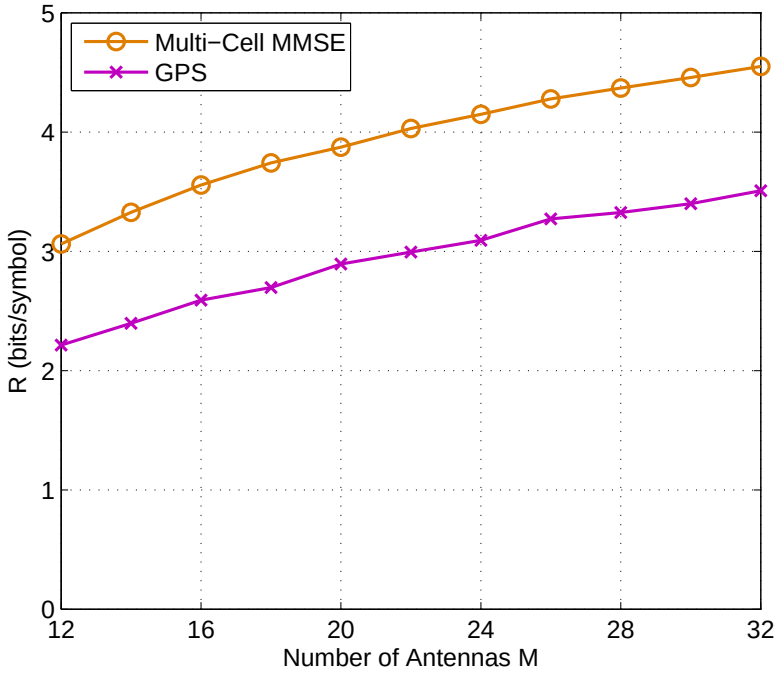

Figure 7.6: Comparison of GPS and multi-cell MMSE with $a = 0.8$ and $b = 0.1a$; R denotes the minimum rate achieved by all users

consider $p_f = 20$ dB and $p_r = 10$ dB. Orthogonal training sequences are collectively used within the 1-st and 2-nd cells. The training sequences used in the 1-st (2-nd) cell are reused in the 3-rd (4-th) cell. Thus, we model a scenario where training sequences are reused. We keep the propagation factors as follows: for all k, $\beta_{jlk} = 1$ if $j = l$, $\beta_{jlk} = a$ if $(j, l) \in \{(1, 2), (2, 1), (3, 4), (4, 3)\}$, and $\beta_{jlk} = b$ for all other values of j and l. "Frequency reuse" is handled semi-quantitatively by adjusting the cross-gains.

Another performance metric of interest is the minimum rate achieved by all users denoted by $R = \min_{jk} R_{jk}$. In Figure 7.5, we plot the performance

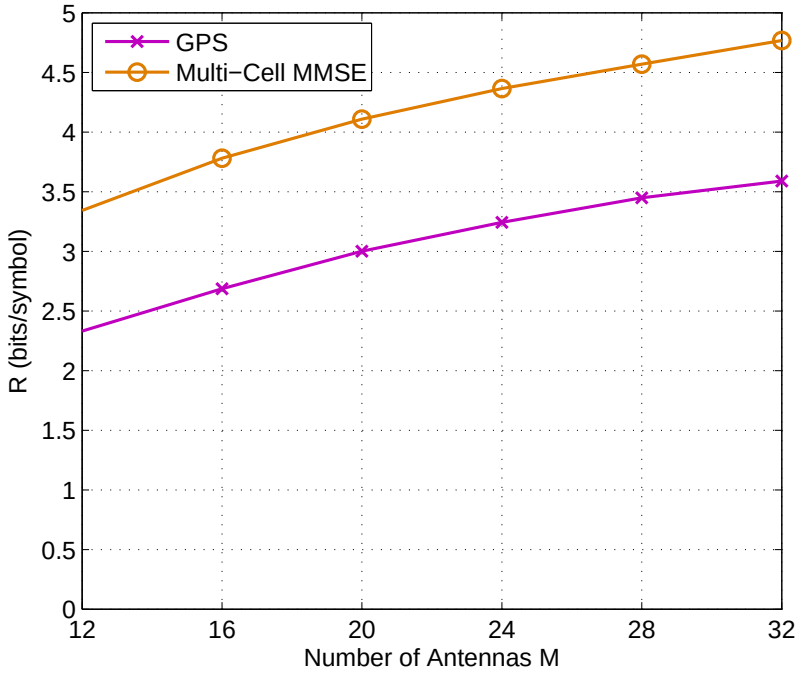

Figure 7.7: Comparison of GPS and multi-cell MMSE with $a = 0.8$ and $b = 0.1a$; Pilots in cells 1 and 3 (2 and 4) are not reused but rotated by 45 degrees

of ZF and multi-cell MMSE precoding methods for different values of a and b. We observe significant advantage of using multi-cell MMSE precoding for wide range of values of a and b. In Figure 7.6, we plot the performance of GPS and multi-cell MMSE precoding methods as a function of the number of antennas M. We also consider the scenario when pilots in cells 1 and 3 (2 and 4) are not reused but rotated by 45 degrees. This comparison is given in Figure 7.7. In both cases, we observe significant advantages in using our multi-cell MMSE precoding. Thus, the multi-cell MMSE scheme is capable of handling different scenarios as it utilizes training sequences for precoder design to mitigate inter-

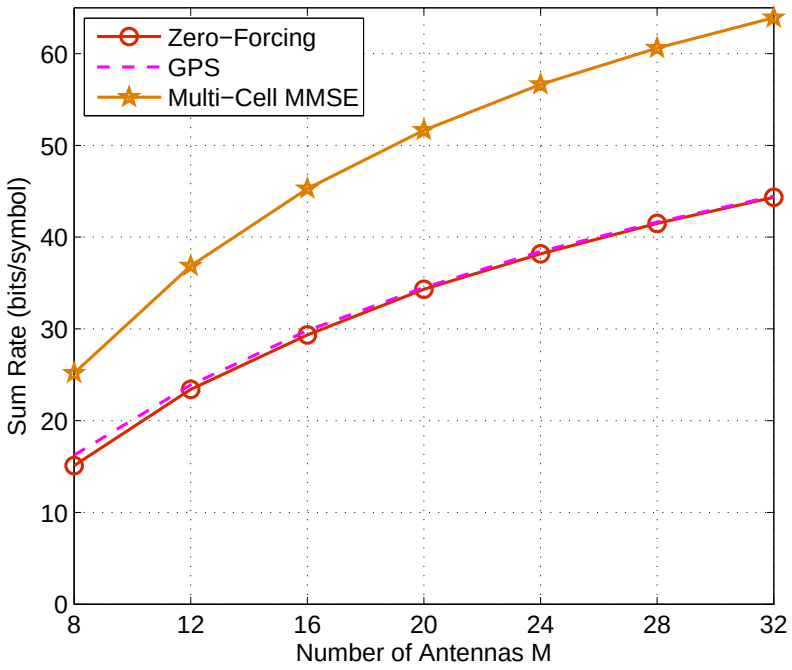

Figure 7.8: Comparison of schemes with $a = 0.8$, $b = 0.1a$, $\tau = 8$; Each cell has $K = 4$ users; Sum rate corresponds to total sum throughput of all cells

cell interference even in the presence of pilot contamination. Note that in both cases the performance of ZF and GPS are almost indistinguishable, and hence we have omitted ZF in these plots.

Finally, we consider a system with reused pilots, $K = 4$ users in every cell and training length of $\tau = 8$. In Figure 7.8, we plot the total sum throughput of different precoding methods as a function of the number of antennas M. In summary, all the numerical results show that the new multi-cell MMSE precoding offers significant performance gains.

Chapter 8

Conclusion

Any communication network predominantly deals with compression and transmission of sources over channels. Hence, it is natural to envisage source and channel aware resource allocation for any communication network. However, this could complicate network design, and therefore, such an approach is desirable only if there are far-reaching consequences. In the context of wireless networks, this is definitely the case due to the following reasons. From a source perspective, the fraction of video content being transferred is steadily growing along with the tremendous growth of overall traffic. From a channel perspective, the interference issue is worsening as a result of decreasing cell-sizes along with aggressive frequency reuse.

In this dissertation, we formulate and study the problem of source and channel aware resource allocation for wireless networks, and solve it in various scenarios. Some of these results have already been published in [63, 64, 67–70]. In Chapter 2, we develop a framework for this by incorporating lossy compression into network resource allocation. Then, we provide a provably optimal layered architecture for performing network optimization with minimal deviation from existing architecture: (i) an application-layer with compression

control, (*ii*) a transport-layer with congestion control, and (*iii*) a network-layer with scheduling. We focus on the compression control problem in this chapter, and develop insights into this problem. In the context of existing networks, our findings suggest a re-design of network architecture with dynamic compression control along with congestion control, and limited exchange of control information across layers. This calls for further study and raises important questions: How to connect the simple source abstraction studied with complicated sources such as video? What are appropriate utility functions in the context of video, and more importantly, in the context of a mixture of different types of traffic?

Next, we focus on two important aspects of network-layer scheduling that have not been sufficiently studied in existing literature. In Chapter 3, we study the problem of developing throughput-optimal network algorithms for cooperative relay networks. These networks are fundamentally different from traditional capacitated and non-cooperative wireless networks, as they require physical-layer coordination. This physical-coordination cannot be abstracted out at the network-layer in terms of bits-in-bits-out models, and thus a stability analysis that incorporates both the physical-layer encoding and the network-layer dynamics is needed. We develop encoding-based queue-architecture for cooperative relay networks. This is a succinct representation needed for generating network-stabilizing algorithms. Using this queue-architecture, we show that throughput-optimal network algorithms can be developed even when the fade-distribution and input queue distributions are unknown. Even though

the queue-architecture we developed brings cooperative schemes closer to reality, there are many challenges yet to be addressed before actual deployment of such schemes. Specifically, future research needs focus on both developing efficient cooperative schemes that require limited coordination, and developing backhaul protocols that provide the coordination required for cooperative schemes.

In Chapter 4, we explore techniques to develop distributed algorithms for the inherently centralized scheduling problem. We successfully extend the current state-of-the-art approach for distributed on-off scheduling to perform distributed rate allocation. These algorithms use local sensing-based and queue-length information, thus, resulting in distributed operation while being aware of its neighborhood. Surprisingly, these algorithms match the throughput performance of optimal centralized algorithms. As known in literature, this algorithmic framework can be combined with network resource allocation, i.e., the arrival rates can be adaptively chosen such that a certain (convex) utility function is maximized as discussed in Chapter 2. Our approach can even be applied to similar problems in other areas, for example, in performing resource allocation in energy networks. However, further research is required in improving delay performance. This has to be addressed for specific classes of network configurations. Additionally, in the context of wireless ad hoc networks, the channel measurement framework introduced motivates further research. First, we need to better understand the feasibility of channel measurement with existing and newly developed radios. This needs development

of good physical-layer architectures that minimize the probability of inaccurate measurement and measurement delay. Further, we need to study the impact of imperfect channel measurement on throughput. We have already made progress along this direction for 802.11 networks in [11], which validates the applicability of channel measurement in practice.

Finally, we make progress in our understanding of maximum achievable data-rates for important classes of interference networks. In Chapter 5, we develop a family of outer bounds for the K-user Gaussian interference channel based on constructing multiple-antenna genie-MAC receivers. This formulation results in an optimization problem that may not be easy to solve in the general case. However, we subsequently show that this family of outer bounds determine the exact sum-capacity of the class of degraded channels, and provide closed-form expression for the sum-capacity. Many intriguing questions stem from this result. Are genie-MAC bounds optimal for other classes of networks? Are there techniques to extend the sum capacity result to non-degraded networks?

In Chapter 6, we study linear precoding for multiple antenna interference channels under channel uncertainty as robust optimization problems. We adopt an alternating optimization approach to solve the original non-convex problems, wherein convex sub-problems are solved at each step. Due to their convexity, the sub-problems can be solved optimally and we enjoy guaranteed convergence. As a result, we develop new schemes that gracefully degrade with increasing channel uncertainty. Based on genie-MACs, we provide computable

outer bounds. There are many open problems that need further study. Even if we focus on linear schemes, as we do in this chapter, efficient algorithms to design optimal precoding matrices do not exist to date. Another open problem is in deriving reasonable bounds on the gap from optimality while restricting to linear schemes. Finally, of course, the open problem of utmost importance is the entire capacity region of these interference networks.

In Chapter 7, we characterize the impact of corrupted channel estimates caused by pilot contamination in time-division-duplex systems. When non-orthogonal training sequences are assigned to users, the precoding matrix used at a (multiple antenna) base station becomes correlated with the channel to users in other cells (referred to as pilot contamination). We show that, in the presence of pilot contamination, rates achieved by users saturate with the number of base station antennas. We conclude that appropriate frequency/time reuse techniques have to be employed to overcome this saturation effect. We develop a multi-cell MMSE-based precoding that depends on the set of training sequences assigned to the users. We obtain this precoding as the solution to an optimization problem whose objective function consists of two parts: (i) the mean-square error of signals received at the users in the same cell, and (ii) the mean-square interference caused at the users in other cells. Through numerical results, we show that our method outperforms popular single-cell precoding methods. Thus, our results suggest that efficient interference management provide significant gains in practice.

Appendices

Appendix A

Proofs of Lemmas for Chapter 3

A.1 Proof of Lemma 3.1

Consider any static split rule characterized by $a_{\mathbf{f}}^{m,\mathbf{g}}$ and $b_{\mathbf{f}}^{m,\mathbf{g}}$. Note that $a_{\mathbf{f}}^{m,\mathbf{g}}$ is the fraction of time for which packets corresponding to encoding scheme m and state $\mathbf{g}$ is transmitted from the source to the relays when the system is in fading state $\mathbf{f}$. Similarly, $b_{\mathbf{f}}^{m,\mathbf{g}}$ is the fraction of time for which these packets are transmitted from the relays to the destinations. Now, this is a valid static split rule only if the following conditions are satisfied: flow conservation constraint for the source in (3.5), flow conservation constraint for each encoding scheme and state in (3.6), and the time conservation constraint for each fading-state in (3.7). A central controller with the knowledge of the fading distribution can also achieve these rates using static time-division. Hence, these are necessary and sufficient conditions. □

A.2 Proof of Lemma 3.4

Since $\boldsymbol{\lambda}+\epsilon\mathbf{1}$ is in the throughput region, from Lemma 3.1, there exists $\hat{a}_{\mathbf{f}}^{m,\mathbf{g}} \geq 0$ and $\hat{b}_{\mathbf{f}}^{m,\mathbf{g}} \geq 0$ for all $m \in \mathcal{M}$, $\mathbf{g} \in \hat{\mathcal{F}}$ and $\mathbf{f} \in \hat{\mathcal{F}}$ such that

$$\boldsymbol{\lambda}+\epsilon\mathbf{1} = \sum_{m,\mathbf{g},\mathbf{f}} \pi_{\mathbf{f}} \hat{a}_{\mathbf{f}}^{m,\mathbf{g}} \mathbf{r}_m \mathbf{1}_{\{(\mathbf{f},\mathbf{g})\in\mathcal{I}_1\}} \mathbf{1}_{\{(m,\mathbf{g})\in\mathcal{I}\}}, \tag{A.1}$$

$$\sum_{\mathbf{f}\in\hat{\mathcal{F}}} \pi_{\mathbf{f}} \hat{a}_{\mathbf{f}}^{m,\mathbf{g}} \mathbf{1}_{\{(\mathbf{f},\mathbf{g})\in\mathcal{I}_1\}} = \sum_{\mathbf{f}\in\hat{\mathcal{F}}} \pi_{\mathbf{f}} \hat{b}_{\mathbf{f}}^{m,\mathbf{g}} \mathbf{1}_{\{(\mathbf{f},\mathbf{g})\in\mathcal{I}_2\}}, \forall (m,\mathbf{g}) \in \mathcal{I}, \tag{A.2}$$

$$\sum_{m,\mathbf{g}} \hat{a}_{\mathbf{f}}^{m,\mathbf{g}} + \hat{b}_{\mathbf{f}}^{m,\mathbf{g}} \leq 1, \forall \mathbf{f}.$$

Consider $\bar{a}_{\mathbf{f}}^{m,\mathbf{g}}$ and $\bar{b}_{\mathbf{f}}^{m,\mathbf{g}}$ obtained as follows:

$$\bar{a}_{\mathbf{f}}^{m,\mathbf{g}} = 0, \forall (\mathbf{f},\mathbf{g}) \notin \mathcal{I}_1, \forall (m,\mathbf{g}) \notin \mathcal{I}, \tag{A.3}$$

$$\bar{a}_{\mathbf{f}}^{m,\mathbf{g}} = \hat{a}_{\mathbf{f}}^{m,\mathbf{g}}, \text{otherwise, and,}$$

$$\bar{b}_{\mathbf{f}}^{m,\mathbf{g}} = 0, \forall (\mathbf{f},\mathbf{g}) \notin \mathcal{I}_2, \forall (m,\mathbf{g}) \notin \mathcal{I}, \tag{A.4}$$

$$\bar{b}_{\mathbf{f}}^{m,\mathbf{g}} = \hat{b}_{\mathbf{f}}^{m,\mathbf{g}}, \text{otherwise.}$$

Due to the indicator functions in (A.1) and (A.2), and the fact that $0 \leq \bar{a}_{\mathbf{f}}^{m,\mathbf{g}} \leq \hat{a}_{\mathbf{f}}^{m,\mathbf{g}}$ and $0 \leq \bar{b}_{\mathbf{f}}^{m,\mathbf{g}} \leq \hat{b}_{\mathbf{f}}^{m,\mathbf{g}}$, the following constraints are satisfied:

$$\boldsymbol{\lambda}+\epsilon\mathbf{1} \leq \sum_{m,\mathbf{g},\mathbf{f}} \pi_{\mathbf{f}} \bar{a}_{\mathbf{f}}^{m,\mathbf{g}} \mathbf{r}_m, \tag{A.5}$$

$$\sum_{\mathbf{f}\in\hat{\mathcal{F}}} \pi_{\mathbf{f}} \bar{a}_{\mathbf{f}}^{m,\mathbf{g}} \leq \sum_{\mathbf{f}\in\hat{\mathcal{F}}} \pi_{\mathbf{f}} \bar{b}_{\mathbf{f}}^{m,\mathbf{g}}, \forall (m,\mathbf{g}), \tag{A.6}$$

$$\sum_{m,\mathbf{g}} \bar{a}_{\mathbf{f}}^{m,\mathbf{g}} + \bar{b}_{\mathbf{f}}^{m,\mathbf{g}} \leq 1, \forall \mathbf{f}. \tag{A.7}$$

Next, consider the following trivial assignment for $\bar{a}_{\mathbf{f}}^{m,\mathbf{g}}$ and $\bar{b}_{\mathbf{f}}^{m,\mathbf{g}}$ that

satisfy (A.3)-(A.4) and (A.6)-(A.7):

$$\bar{a}_{\mathbf{f}}^{m,\mathbf{g}} = 0, \forall(\mathbf{f},\mathbf{g}),$$
$$\bar{b}_{\mathbf{f}}^{m,\mathbf{g}} = 0, \forall(\mathbf{f},\mathbf{g}) \notin \mathcal{I}_2, \forall(m,\mathbf{g}) \notin \mathcal{I},$$
$$\bar{b}_{\mathbf{f}}^{m,\mathbf{g}} = \frac{1}{|\mathcal{M}||\hat{\mathcal{F}}|}, \text{otherwise.}$$

Since $\epsilon > 0$, there exists $\phi < 1$ such that $\boldsymbol{\lambda} \leq \phi(\boldsymbol{\lambda} + \epsilon\mathbf{1})$. Now, since (A.3)-(A.4) and (A.6)-(A.7) are linear constraints, the following convex combination will also satisfy these constraints:

$$a_{\mathbf{f}}^{m,\mathbf{g}} = 0, \forall(\mathbf{f},\mathbf{g}) \notin \mathcal{I}_1, \forall(m,\mathbf{g}) \notin \mathcal{I},$$
$$a_{\mathbf{f}}^{m,\mathbf{g}} = \frac{\phi+1}{2}\hat{a}_{\mathbf{f}}^{m,\mathbf{g}}, \text{otherwise,}$$
$$b_{\mathbf{f}}^{m,\mathbf{g}} = 0, \forall(\mathbf{f},\mathbf{g}) \notin \mathcal{I}_2, \forall(m,\mathbf{g}) \notin \mathcal{I}.$$
$$b_{\mathbf{f}}^{m,\mathbf{g}} = \frac{\phi+1}{2}\hat{b}_{\mathbf{f}}^{m,\mathbf{g}} + \frac{1-\phi}{2}\frac{1}{|\mathcal{M}||\hat{\mathcal{F}}|}, \text{otherwise.}$$

From (A.5), we get

$$\lambda_k < \frac{\phi+1}{2}(\lambda_k + \epsilon) \leq \sum_{m,\mathbf{g},\mathbf{f}} \pi_{\mathbf{f}} a_{\mathbf{f}}^{m,\mathbf{g}} r_m^k, \forall k.$$

From (A.2), we get

$$\begin{aligned}
\sum_{\mathbf{f}\in\hat{\mathcal{F}}} \pi_{\mathbf{f}} a_{\mathbf{f}}^{m,\mathbf{g}} &= \frac{\phi+1}{2}\sum_{\mathbf{f}\in\hat{\mathcal{F}}} \pi_{\mathbf{f}} \hat{a}_{\mathbf{f}}^{m,\mathbf{g}} \mathbf{1}_{\{(\mathbf{f},\mathbf{g})\in\mathcal{I}_1\}} \mathbf{1}_{\{(m,\mathbf{g})\in\mathcal{I}\}} \\
&= \frac{\phi+1}{2}\sum_{\mathbf{f}\in\hat{\mathcal{F}}} \pi_{\mathbf{f}} \hat{b}_{\mathbf{f}}^{m,\mathbf{g}} \mathbf{1}_{\{(\mathbf{f},\mathbf{g})\in\mathcal{I}_2\}} \mathbf{1}_{\{(m,\mathbf{g})\in\mathcal{I}\}} \\
&= \sum_{\mathbf{f}\in\hat{\mathcal{F}}} \pi_{\mathbf{f}} b_{\mathbf{f}}^{m,\mathbf{g}} - \frac{1-\phi}{2}\frac{1}{|\mathcal{M}||\hat{\mathcal{F}}|}\sum_{\mathbf{f}\in\hat{\mathcal{F}}} \pi_{\mathbf{f}} \mathbf{1}_{\{(\mathbf{f},\mathbf{g})\in\mathcal{I}_2\}} \mathbf{1}_{\{(m,\mathbf{g})\in\mathcal{I}\}} \\
&< \sum_{\mathbf{f}\in\hat{\mathcal{F}}} \pi_{\mathbf{f}} b_{\mathbf{f}}^{m,\mathbf{g}}, \forall(m,\mathbf{g}).
\end{aligned}$$

In the last step, the strict inequality requires that $\sum_{\mathbf{f}\in\hat{\mathcal{F}}} \pi_{\mathbf{f}} \mathbf{1}_{\{(\mathbf{f},\mathbf{g})\in\mathcal{I}_2\}} \mathbf{1}_{\{(m,\mathbf{g})\in\mathcal{I}\}} > 0$. If this is not the case for any given $(m, \mathbf{g})$, those virtual queues can be neglected from the analysis as it does not affect the throughput region. Thus, we have constructed $a_{\mathbf{f}}^{m,\mathbf{g}}$ and $b_{\mathbf{f}}^{m,\mathbf{g}}$ that satisfy all the conditions stated in the lemma. $\square$

Appendix B

Proofs of Lemmas for Chapter 4

The proofs given in this Appendix are generalizations of proofs in [57, 58] to multi-state framework. Since we cannot directly apply existing results, we provide the complete proofs here.

B.1 Proof of Lemma 4.4

The steps involved are the following. First, we prove that, for any fixed $\boldsymbol{\lambda} \in \mathbb{R}^n_+$, the objective function $F(\mathbf{v}, \boldsymbol{\lambda})$ is *strictly* concave in $\mathbf{v}$. Next, we show that for any fixed $\boldsymbol{\lambda} \in \mathcal{R}^o_c$, the optimal value $\mathbf{v}^*$ lies inside a compact subset of $\mathbb{R}^n$. These two statements show the existence of a unique solution that is finite. This along with certain necessary condition for optimality completes the proof.

For notational simplicity, we denote $F(\mathbf{v}, \boldsymbol{\lambda})$ by $F(\mathbf{v})$ and the normalization constant or partition function by $Z(\mathbf{v}) := \sum_{\mathbf{r} \in \mathcal{R}} \exp(\mathbf{r} \cdot \mathbf{v})$. Using calculus, it is straightforward to obtain the gradient (first-order partial derivatives) and the Hessian (second-order partial derivatives) of $F(\mathbf{v})$ in the following

form:

$$\begin{aligned} \nabla F(\mathbf{v}) &= \boldsymbol{\lambda} - \mathbb{E}_{\pi_{\mathbf{v}}}[\mathbf{r}] \\ &= \boldsymbol{\lambda} - \mathbf{s}_{\mathbf{v}}; \end{aligned} \tag{B.1}$$

$$H(F(\mathbf{v})) = -\left(\mathbb{E}_{\pi_{\mathbf{v}}}\left[\mathbf{r}\mathbf{r}^T\right] - \mathbb{E}_{\pi_{\mathbf{v}}}\left[\mathbf{r}\right]\mathbb{E}_{\pi_{\mathbf{v}}}\left[\mathbf{r}^T\right]\right). \tag{B.2}$$

Here, $\mathbf{s}_{\mathbf{v}}$ in (B.1) is the offered service rate vector given by (4.13), and $\mathbb{E}_{\pi_{\mathbf{v}}}[\Phi] := \sum_{\mathbf{r}\in\mathcal{R}} \pi_{\mathbf{v}}(\mathbf{r})\Phi$ for any matrix, vector or scalar Φ.

In order to establish that $F(\mathbf{v})$ is *strictly* concave in $\mathbf{v}$, we show that the Hessian H is negative definite, i.e., for any non-zero $\boldsymbol{\eta} \in \mathcal{R}^n$, $\boldsymbol{\eta}^T H \boldsymbol{\eta} < 0$. Since H is the negative of a covariance matrix, it is clear that H is negative semi-definite, i.e., from (B.2),

$$\begin{aligned} \boldsymbol{\eta}^T H \boldsymbol{\eta} &= -\mathbb{E}_{\pi_{\mathbf{v}}}\left[\boldsymbol{\eta}^T(\mathbf{r} - \mathbb{E}_{\pi_{\mathbf{v}}}[\mathbf{r}])(\mathbf{r} - \mathbb{E}_{\pi_{\mathbf{v}}}[\mathbf{r}])^T\boldsymbol{\eta}\right] \\ &= -\mathbb{E}_{\pi_{\mathbf{v}}}\left[\left(\boldsymbol{\eta}^T(\mathbf{r} - \mathbb{E}_{\pi_{\mathbf{v}}}[\mathbf{r}])\right)^2\right] \leq 0. \end{aligned} \tag{B.3}$$

We next prove that the Hessian H is negative definite by contradiction. Consider a fixed $\mathbf{v}$. Suppose that there exists $\boldsymbol{\eta} \neq 0$ such that $\boldsymbol{\eta}^T H \boldsymbol{\eta} = 0$. Then, from (B.3), it follows that the random variable $\boldsymbol{\eta}^T(\mathbf{r} - \mathbb{E}_{\pi_{\mathbf{v}}}[\mathbf{r}])$ is zero with probability 1. For any fixed $\mathbf{v}$, all feasible states have non-zero probability. In particular, $\pi_{\mathbf{v}}(\mathbf{0}) > 0$ and $\pi_{\mathbf{v}}(c_i\mathbf{e}_i) > 0$ for all $i \in \mathcal{L}$. Therefore, the random variable must evaluate to zero at $\mathbf{r} = \mathbf{0}$ and $\mathbf{r} = c_i\mathbf{e}_i$, i.e.,

$$-\boldsymbol{\eta}^T\mathbb{E}_{\pi_{\mathbf{v}}}[\mathbf{r}] = \eta_i c_i - \boldsymbol{\eta}^T\mathbb{E}_{\pi_{\mathbf{v}}}[\mathbf{r}] = 0,$$

which implies $\boldsymbol{\eta} = 0$. This provides a contradiction and establishes that the Hessian H is negative definite.

Next, we prove that the optimal value $\mathbf{v}^*$ belongs to a compact set. Let $\boldsymbol{\lambda} + \delta\bar{K}\mathbf{1} \in \mathcal{R}_c$ for some $0 < \delta < 1$. Note that for any $\boldsymbol{\lambda} \in \mathcal{R}_c^o$ there exists such a δ. Consider a $\mathbf{v} \in \mathbb{R}^n$. Define $v_{\min} = \min_i v_i$, $l = \arg\min_i v_i$, and $v_{\max} = \max_i v_i$. Let

$$\hat{\boldsymbol{\lambda}} = \boldsymbol{\lambda} - \min(\delta\bar{K}, \lambda_{\min})\mathbf{I}(v_{\min} < 0)\mathbf{e}_l.$$

Clearly, $\hat{\boldsymbol{\lambda}} + \min(\delta\bar{K}, \lambda_{\min})\mathbf{1} \in \mathcal{R}_c$, and hence, there exists a distribution μ on $\mathcal{R}$ such that $\hat{\boldsymbol{\lambda}} + \min(\delta\bar{K}, \lambda_{\min}) = \mathbb{E}_\mu[\mathbf{r}]$. Since $\hat{\boldsymbol{\lambda}} \leq \bar{K}\mathbf{1}$, we have

$$\hat{\boldsymbol{\lambda}} \leq \frac{\hat{\boldsymbol{\lambda}} + \min(\delta\bar{K}, \lambda_{\min})}{1 + \min(\delta, \lambda_{\min}/\bar{K})} = \sum_{\mathbf{r}\in\mathcal{R}} \frac{\mu(\mathbf{r})\mathbf{r}}{1 + \min(\delta, \lambda_{\min}/\bar{K})} \tag{B.4}$$

and

$$\begin{aligned} \sum_{\mathbf{r}\in\mathcal{R}} \frac{\mu(\mathbf{r})}{1 + \min(\delta, \lambda_{\min}/\bar{K})} &= \frac{1}{1 + \min(\delta, \lambda_{\min}/\bar{K})} \\ &< 1 - \frac{\min(\delta, \lambda_{\min}/\bar{K})}{2}. \end{aligned} \tag{B.5}$$

From (B.4), (B.5) and the fact that $\mathbf{0}, c_i\mathbf{e}_i \in \mathcal{R}$, it follows that there exists a non-negative measure $\hat{\mu}$ on $\mathcal{R}$ such that $\hat{\boldsymbol{\lambda}} = \sum_{\mathbf{r}\in\mathcal{R}} \hat{\mu}(\mathbf{r})\mathbf{r}$ with $\sum_{\mathbf{r}\in\mathcal{R}} \hat{\mu}(\mathbf{r}) = 1 - 0.5\min(\delta, \lambda_{\min}/\bar{K})$. Now, define a distribution

$$\tilde{\mu}(\mathbf{r}) = \begin{cases} \hat{\mu}(c_l\mathbf{e}_l) + \frac{\min(\delta,\lambda_{\min}/\bar{K})}{4}\mathbf{I}(v_{\min} < 0), & \text{if } \mathbf{r} = c_l\mathbf{e}_l, \\ \hat{\mu}(\mathbf{0}) + \frac{\min(\delta,\lambda_{\min}/\bar{K})}{4}\left(2 - \mathbf{I}(v_{\min} < 0)\right), & \text{if } \mathbf{r} = \mathbf{0}, \\ \hat{\mu}(\mathbf{r}), & \text{otherwise.} \end{cases}$$

Define $\tilde{\boldsymbol{\lambda}} = \mathbb{E}_{\tilde{\mu}}[\mathbf{r}]$. Now, we have

$$\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda} - \left(1 - \frac{c_l}{4\bar{K}}\right)\min(\delta\bar{K}, \lambda_{\min})\mathbf{I}(v_{\min} < 0)\mathbf{e}_l.$$

Clearly, $\boldsymbol{\lambda} \cdot \mathbf{v} \leq \tilde{\boldsymbol{\lambda}} \cdot \mathbf{v}$. Substituting these inequalities in (4.15), we obtain

$$\begin{aligned}
F(\mathbf{v}) &= \boldsymbol{\lambda} \cdot \mathbf{v} - \log Z(\mathbf{v}) \\
&\leq \tilde{\boldsymbol{\lambda}} \cdot \mathbf{v} - \log Z(\mathbf{v}) \\
&= \sum_{\mathbf{r} \in \mathcal{R}} \tilde{\mu}(\mathbf{r}) \mathbf{r} \cdot \mathbf{v} - \log Z(\mathbf{v}) = \sum_{\mathbf{r} \in \mathcal{R}} \tilde{\mu}(\mathbf{r}) \log \frac{\exp(\mathbf{r} \cdot \mathbf{v})}{Z(\mathbf{v})} \\
&\overset{(a)}{\leq} \min \left(\tilde{\mu}(c_l \mathbf{e}_l) \log \frac{\exp(c_l \mathbf{e}_l \cdot \mathbf{v})}{Z(\mathbf{v})}, \tilde{\mu}(\mathbf{0}) \log \frac{\exp(\mathbf{0} \cdot \mathbf{v})}{Z(\mathbf{v})} \right) \\
&\overset{(b)}{\leq} \min \left(\frac{\min(\delta, \lambda_{\min}/\bar{K}) \mathbf{I}(v_{\min} < 0)}{4} \log \frac{\exp(\bar{K} v_{\min})}{1}, \right. \\
&\qquad \left. \frac{\min(\delta, \lambda_{\min}/\bar{K})}{4} \log \frac{1}{\exp(\underline{K} v_{\max})} \right). \qquad \text{(B.6)}
\end{aligned}$$

Here, (a) follows from $\exp(\mathbf{r} \cdot \mathbf{v}) \leq Z(\mathbf{v})$ for any $\mathbf{r} \in \mathcal{R}$, and (b) follows from $\underline{K} \leq c_i \leq \bar{K}$ for any $i \in \mathcal{L}$. Let $\mathbf{v}^* = \sup_{\mathbf{v} \in \mathbb{R}^n} F(\mathbf{v})$. Then, by definition, $F(\mathbf{v}^*) \geq F(\mathbf{0}) = -\log |\mathcal{R}|$. From (B.6), we obtain the bounds

$$v^*_{\max} \leq \frac{4 \log |\mathcal{R}|}{\underline{K} \min(\delta, \lambda_{\min}/\bar{K})} \qquad \text{(B.7)}$$

and

$$v^*_{\min} \geq -\frac{4 \log |\mathcal{R}|}{\bar{K} \min(\delta, \lambda_{\min}/\bar{K})}. \qquad \text{(B.8)}$$

Thus, there exists a unique solution which is finite. Finally, the necessary condition in (B.1) for optimality completes the proof. □

B.2 Proof of Lemma 4.5

The first part of the proof follows directly from Lemma 4.4. The second part also follows from the proof of Lemma 4.4 as explained next. In the proof,

replace $\boldsymbol{\lambda}$ with $\boldsymbol{\lambda} + \frac{\epsilon}{4}\mathbf{1}$ and choose $\delta = \frac{\epsilon}{4\bar{K}}$. Now from (B.7), (B.8) and (4.6), we obtain

$$\|\mathbf{v}^*\|_\infty \leq \frac{4n\bar{K}\log\left\lceil 2\bar{K}/\epsilon\right\rceil}{\min\left(\frac{\epsilon}{4}, \lambda_{\min}\right)}\frac{1}{\underline{K}}. \tag{B.9}$$

This follows from $\underline{K} \leq \bar{K}$. If $\epsilon \leq 4\lambda_{\min}$, then (B.9) simplifies to (4.23). □

B.3 Proof of Lemma 4.6

Consider the matrix $\hat{P} = \exp(P - I)$. It is fairly straightforward to verify that $\hat{P}$ corresponds the probability transmission matrix of a reversible Markov chain with the same stationary distribution $\boldsymbol{\pi}_{\mathbf{v}}$. Now, the steps involved to complete the proof are the following. We need to obtain a lower bound on the conductance associated with $\hat{P}$ and apply Result 4.2. Then, we can apply Result 4.1 to $\hat{P}$ at $\tau = \lfloor At \rfloor$.

From (4.14), $\pi_{\mathbf{v}}(\mathbf{r}) = (\exp(\mathbf{r}\cdot\mathbf{v}))/Z(\mathbf{v})$, where the partition function

$$Z(\mathbf{v}) = \sum_{\mathbf{r}\in\mathcal{R}} \exp(\mathbf{r}\cdot\mathbf{v}).$$

From (4.6), it is clear that $Z(\mathbf{v}) \leq \left\lceil 2\bar{K}/\epsilon\right\rceil^n \exp(\bar{K}n\|\mathbf{v}\|_\infty)$. In addition, $\exp(\mathbf{r}\cdot\mathbf{v}) \geq \exp(-\bar{K}n\|\mathbf{v}\|_\infty)$. Therefore, for all $\mathbf{r} \in \mathcal{R}$,

$$\pi_{\mathbf{v}}(\mathbf{r}) \geq \frac{\exp(-2\bar{K}n\|\mathbf{v}\|_\infty)}{\left\lceil 2\bar{K}/\epsilon\right\rceil^n}. \tag{B.10}$$

Consider two states that differ in one dimension, i.e., $\mathbf{r}, \hat{\mathbf{r}} \in \mathcal{R}$, $\|\mathbf{r}-\hat{\mathbf{r}}\|_0 = 1$, then the transition probability $\hat{P}(\mathbf{r}, \hat{\mathbf{r}})$ is lower bounded by the product of the probability that a Poisson random variable with parameter 1 is one

and $P(\mathbf{r}, \hat{\mathbf{r}})$. This follows from the fact that these two (independent) events together contribute to the transition probability $\hat{P}(\mathbf{r}, \hat{\mathbf{r}})$. Hence,

$$\begin{aligned} \hat{P}(\mathbf{r}, \hat{\mathbf{r}}) &\geq e^{-1} P(\mathbf{r}, \hat{\mathbf{r}}) \\ &= e^{-1} \frac{f(\mathbf{r}, \hat{\mathbf{r}})}{A} \\ &\geq \frac{\exp(-2\bar{K}\|\mathbf{v}\|_\infty)}{ne}, \end{aligned}$$

where $f(\mathbf{r}, \hat{\mathbf{r}})$ is given by (4.24) and $A = n \exp(\bar{K}\|\mathbf{v}\|_\infty)$. To lower bound conductance in (4.4), the following observation can be used. If both $\mathcal{S}$ and $\mathcal{S}^c$ are non-empty, then there is at least one state in $\mathcal{S}$ and another state in $\mathcal{S}^c$ that differ in one dimension alone. This follows from the fact that the state-space is connected through these one dimensional transitions alone. Applying this, we obtain

$$\Phi \geq \frac{\exp(-2\bar{K}(n+1)\|\mathbf{v}\|_\infty)}{ne \left\lceil 2\bar{K}/\epsilon \right\rceil^n}. \tag{B.11}$$

Using (4.5), and substituting (B.11), (B.10) in (4.3), we have the required result $\|\boldsymbol{\mu}(t) - \boldsymbol{\pi}_{\mathbf{v}}\|_{TV} \leq \rho_1$ if

$$t = \exp\left(\Theta\left(n\|\mathbf{v}\|_\infty + n \log \frac{1}{\epsilon}\right)\right) \log \frac{1}{\rho_1}.$$

This completes the proof. □

B.4 Proof of Lemma 4.7

In the proof, we suppress l in the notation, denote $\mathbf{s_v}$ by $\mathbf{s}$, and denote $\pi_{\mathbf{v}}$ by π. From triangle inequality and linearity of expectations, we have

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{\lambda}}(l)-\boldsymbol{\lambda}\right\|_1\right]+\mathbb{E}\left[\|\hat{\mathbf{s}}(l)-\mathbf{s_v}(l)\|_1\right]\leq\sum_{i=1}^{n}\mathbb{E}\left[|\hat{\lambda_i}-\lambda_i|\right]+\sum_{i=1}^{n}\mathbb{E}\left[|\hat{s}_i-\mathbb{E}[\hat{s}_i]|\right]+\sum_{i=1}^{n}|\mathbb{E}[\hat{s}_i]-s_i|. \tag{B.12}$$

Now, we focus on i-th link and upper bound each of the three terms on the RHS of (B.12) corresponding to this link separately by $\rho_2/3n$.

For bounding the first term in (B.12), denote the arrivals over integral times as $\{\xi_k\}_{k=1}^{T}$. From our assumption on arrival processes, these are i.i.d. random variables with variance at most K^2. Hence,

$$\begin{aligned}\mathbb{E}\left[|\hat{\lambda_i}-\lambda_i|\right] &\leq \left(\mathbb{E}\left[(\hat{\lambda_i}-\lambda_i)^2\right]\right)^{\frac{1}{2}}\\ &= \left(\mathbb{E}\left[\left(\frac{1}{T}\sum_{k=1}^{T}\xi_k-\lambda_i\right)^2\right]\right)^{\frac{1}{2}}\\ &\leq \frac{K}{\sqrt{T}}.\end{aligned} \tag{B.13}$$

Next, we consider the expected offered service rate under distribution $\boldsymbol{\mu}(t)$, where $\boldsymbol{\mu}(t)$ denotes the distribution over $\mathcal{R}$ given by the algorithm at time t. From (4.13), we have

$$\begin{aligned}|\mathbb{E}_{\mu(t)}[r_i]-s_i| &= |\mathbb{E}_{\mu(t)}[r_i]-\mathbb{E}_{\pi}[r_i]|\\ &\leq 2\bar{K}\|\mu(t)-\pi\|_{TV}.\end{aligned} \tag{B.14}$$

If we look at two times z and y such that $0 \leq z \leq y$, then

$$\begin{aligned}\mathbb{E}[r_i(z)r_i(y)] &= \mathbb{E}[r_i(z)\mathbb{E}[r_i(y)|r_i(z)]] \\ &\leq \mathbb{E}[r_i(z)] \max_{\beta \in \mathcal{R}_i} \mathbb{E}[r_i(y)|r_i(z) = \beta]. \end{aligned} \tag{B.15}$$

We use (B.14) and (B.15) along with Lemma 4.6 to obtain bounds on the last two terms in (B.12). Let $B(\rho_1)$ be large enough time such that it satisfies (4.26).

For the second term in (B.12), using (4.8), we have

$$\begin{aligned}&\left(\mathbb{E}\left[|\hat{s}_i - \mathbb{E}[\hat{s}_i]|\right]\right)^2 \leq \mathbb{E}\left[(\hat{s}_i - \mathbb{E}[\hat{s}_i])^2\right] \\ &= \mathbb{E}\left[(\hat{s}_i)^2\right] - (\mathbb{E}[\hat{s}_i])^2 \\ &= \mathbb{E}\left[\left(\frac{1}{T}\int_0^T r_i(z)dz\right)^2\right] - \left(\mathbb{E}\left[\frac{1}{T}\int_0^T r_i(z)dz\right]\right)^2 \\ &= \frac{1}{T^2}\int_0^T \int_0^T \left(\mathbb{E}\left[r_i(z)r_i(y)\right] - \mathbb{E}\left[r_i(z)\right]\mathbb{E}\left[r_i(y)\right]\right) dydz \\ &= \frac{2}{T^2}\int_0^T \int_z^T \left(\mathbb{E}\left[r_i(z)r_i(y)\right] - \mathbb{E}\left[r_i(z)\right]\mathbb{E}\left[r_i(y)\right]\right) dydz \\ &\leq \frac{2}{T^2}\int_0^T \mathbb{E}[r_i(z)]\hat{I}dz, \end{aligned} \tag{B.16}$$

where the inner integral

$$\hat{I} = \int_z^T \left(\max_{\beta \in \mathcal{R}_i} \mathbb{E}[r_i(y)|r_i(z) = \beta] - \mathbb{E}\left[r_i(y)\right]\right) dy.$$

Here, we used (B.15). Now, from (B.14) and Lemma 4.6 on mixing time, both $\max_{\beta \in \mathcal{R}_i} \mathbb{E}[r_i(y)|r_i(z) = \beta]$ and $\mathbb{E}\left[r_i(y)\right]$ are *close* to s_i by total variation ρ_1

each if $y \geq z + B(\rho_1)$. Formally, we bound $\hat{I}$ as follows:

$$\begin{aligned} \hat{I} &\leq \int_z^{z+B(\rho_1)} \bar{K} dy + \int_{z+B(\rho_1)}^{T} 4\rho_1 \bar{K} dy \\ &\leq B(\rho_1)\bar{K} + 4\rho_1 \bar{K} T. \end{aligned} \tag{B.17}$$

Substituting (B.17)in (B.16), we obtain

$$\begin{aligned} \mathbb{E}\left[|\hat{s}_i - \mathbb{E}[\hat{s}_i]|\right] &\leq \left(\frac{2}{T^2}\int_0^T \mathbb{E}[r_i(z)](B(\rho_1)\bar{K} + 4\rho_1\bar{K}T)dz\right)^{\frac{1}{2}} \\ &\leq \left(\frac{2}{T}\bar{K}^2 B(\rho_1) + 8\bar{K}^2\rho_1\right)^{\frac{1}{2}}, \end{aligned} \tag{B.18}$$

where we used $\mathbb{E}[r_i(z)] \leq \bar{K}$.

For the third term, from (4.8) and (B.14) and using techniques applied above, we obtain

$$\begin{aligned} |\mathbb{E}[\hat{s}_i] - s_i| &= \left|\frac{1}{T}\int_0^T \mathbb{E}\left[r_i(z)\right] dz - s_i\right| \\ &\leq \frac{\bar{K}B(\rho_1)}{T} + 2\bar{K}\rho_1. \end{aligned} \tag{B.19}$$

With $\rho_1 = \rho_2^2/(144n^2\bar{K}^2)$ and the choice of

$$T = \exp\left(\Theta\left(n\|\mathbf{v}\|_\infty + n\log\frac{1}{\epsilon}\right)\right)\frac{1}{\rho_2},$$

it is fairly straightforward to see that RHS of (B.13), (B.18) and (B.19) can be made smaller than $\rho_2/3n$. This completes the proof of Lemma 4.7. □

B.5 Proof of Lemma 4.8

For simplicity, we denote $\mathbf{v}(\tau_l)$ by $\mathbf{v}_l$. Define $G(\mathbf{v}) := F_\epsilon(\mathbf{v}) - \|\mathbf{v} - \mathbf{v}^*\|_2^2$. Let $[\boldsymbol{\theta}]_D$ denote component-wise $[\theta_i]_D$. This function has the following monotone property.

Lemma B.1. *Consider any* $\mathbf{v} \in [-D, D]^n$, $\Delta\mathbf{v} \in [-1, 1]^n$. *Then,* $G([\mathbf{v} + \Delta\mathbf{v}]_D) \geq G(\mathbf{v} + \Delta\mathbf{v})$. *Also,* $0 \geq G(\mathbf{v}) \geq -7nD^2$.

Proof. Please see Section B.5.1. □

Let the error term in the l-th time interval be

$$\mathbf{e}_l = (\hat{\boldsymbol{\lambda}}(l) - \hat{\mathbf{s}}(l)) - (\boldsymbol{\lambda} - \mathbf{s}_{\mathbf{v}_l})$$

and $\hat{\mathbf{e}}_l = \alpha(\nabla F_\epsilon(\mathbf{v}_l) + \mathbf{e}_l)$. From Lemma 4.5, the update equation in (4.9) can be written as $\mathbf{v}_{l+1} = \mathbf{v}_l + \hat{\mathbf{e}}_l$. We have $\nabla F_\epsilon(\mathbf{v}_l) \in [-\bar{K}, \bar{K}]^n$, $\mathbf{e}_l \in [-\bar{K}$ and $\mathbf{v}_l, \mathbf{v}^* \in [-D, D]^n$. Therefore, $\|\hat{\mathbf{e}}_l\|_\infty \leq \alpha(2\bar{K} + K) \leq 1$. From Lemma B.1 and Taylor's expansion, we obtain

$$\begin{aligned} G(\mathbf{v}_{l+1}) &= G([\mathbf{v}_l + \hat{\mathbf{e}}_l]_D) \\ &\geq G(\mathbf{v}_l + \hat{\mathbf{e}}_l) \\ &= F_\epsilon(\mathbf{v}_l + \hat{\mathbf{e}}_l) - \|\mathbf{v}_l + \hat{\mathbf{e}}_l - \mathbf{v}^*\|_2^2 \\ &= G(\mathbf{v}_l) + \nabla F_\epsilon(\mathbf{v}_l) \cdot \hat{\mathbf{e}}_l + \frac{1}{2}\hat{\mathbf{e}}_l H \hat{\mathbf{e}}_l \\ &\quad -\|\hat{\mathbf{e}}_l\|_2^2 - 2(\mathbf{v}_l - \mathbf{v}^*) \cdot \hat{\mathbf{e}}_l, \end{aligned} \tag{B.20}$$

where H is the Hessian of $F_\epsilon(\cdot)$ evaluated at some $\tilde{\mathbf{v}}$ around $\mathbf{v}_l$. The elements of the matrix H belong to $[-\bar{K}^2, \bar{K}^2]$, $\mathbf{e}_l \in [-\bar{K}, K]^n$, $\nabla F_\epsilon(\mathbf{v}_l) \in [-\bar{K}, \bar{K}]^n$ and $\mathbf{v}_l, \mathbf{v}^* \in [-D, D]^n$. Therefore, $\|\hat{\mathbf{e}}_l\|_\infty \leq \alpha(2\bar{K} + K)$. Using these, we have

$$\frac{1}{2}\hat{\mathbf{e}}_l H \hat{\mathbf{e}}_l - \|\hat{\mathbf{e}}_l\|_2^2 \geq -\alpha^2 c,$$

where $c = (2\bar{K} + K)^2 \left(\frac{\bar{K}^2 n^2}{2} + n \right)$. Since $F_\epsilon(\mathbf{v})$ is concave with optimum $\mathbf{v}^*$,

$$F_\epsilon(\mathbf{v}^*) \leq F_\epsilon(\mathbf{v}_l) + \nabla F_\epsilon(\mathbf{v}_l) \cdot (\mathbf{v}_l - \mathbf{v}^*).$$

It follows that $\nabla F_\epsilon(\mathbf{v}_l) \cdot (\mathbf{v}_l - \mathbf{v}^*) \geq 0$. Applying these to (B.20), we obtain

$$\begin{aligned} G(\mathbf{v}_{l+1}) &\geq G(\mathbf{v}_l) + \alpha \|\nabla F_\epsilon(\mathbf{v}_l)\|_2^2 + \alpha \nabla F_\epsilon(\mathbf{v}_l) \cdot \mathbf{e}_l - \alpha^2 c \\ &\quad -2\alpha(\mathbf{v}_l - \mathbf{v}^*) \cdot \nabla F_\epsilon(\mathbf{v}_l) - 2\alpha(\mathbf{v}_l - \mathbf{v}^*) \cdot \mathbf{e}_l, \\ &\geq G(\mathbf{v}_l) + \alpha \|\nabla F_\epsilon(\mathbf{v}_l)\|_2^2 - \alpha \bar{K} \|\mathbf{e}_l\|_1 - \alpha^2 c \\ &\quad -4\alpha D \|\mathbf{e}_l\|_1, \\ &\geq G(\mathbf{v}_l) + \alpha \|\nabla F_\epsilon(\mathbf{v}_l)\|_2^2 - 5\alpha D \|\mathbf{e}_l\|_1 - \alpha^2 c. \end{aligned}$$

Here, we used $\bar{K} \leq D$.

Next, performing telescopic sum and then using $G(\mathbf{v}_1) \geq -7nD^2$ from Lemma B.1, we obtain,

$$\begin{aligned} G(\mathbf{v}_{N+1}) &= \sum_{l=1}^{N} (G(\mathbf{v}_{l+1}) - G(\mathbf{v}_l)) + G(\mathbf{v}_1) \\ &\geq \alpha \sum_{l=1}^{N} \|\nabla F_\epsilon(\mathbf{v}_l)\|_2^2 - 5\alpha D \sum_{l=1}^{N} \|\mathbf{e}_l\|_1 \\ &\quad -\alpha^2 cN - 7nD^2. \end{aligned}$$

Since $G(\mathbf{v}_{N+1}) \leq 0$, and then applying (4.27), we get

$$\begin{aligned} \frac{1}{N} \sum_{l=1}^{N} \|\nabla F_\epsilon(\mathbf{v}_l)\|_2^2 &\leq \frac{5D}{N} \sum_{l=1}^{N} \|\mathbf{e}_l\|_1 + \alpha c + \frac{7nD^2}{\alpha N} \\ &\leq 5D\rho_2 + \alpha c + \frac{7nD^2}{\alpha N}. \end{aligned} \tag{B.21}$$

Applying Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
\frac{1}{N}\sum_{l=1}^{N}\mathbb{E}\left[\nabla F_\epsilon(\mathbf{v}_l)\right] &\leq \frac{1}{N}\sum_{l=1}^{N}\mathbb{E}\left[\|\nabla F_\epsilon(\mathbf{v}_l)\|_2\right]\mathbf{1} \\
&\leq \sqrt{\frac{1}{N}\sum_{l=1}^{N}\left(\mathbb{E}\left[\|\nabla F_\epsilon(\mathbf{v}_l)\|_2\right]\right)^2}\mathbf{1} \\
&\leq \sqrt{\frac{1}{N}\sum_{l=1}^{N}\mathbb{E}\left[\|\nabla F_\epsilon(\mathbf{v}_l)\|_2^2\right]}\mathbf{1} \\
&\leq \sqrt{5D\rho_2+\alpha c+\frac{7nD^2}{\alpha N}}\mathbf{1}. \qquad \text{(B.22)}
\end{aligned}$$

Next, we look at the average of the empirical service rates over N update intervals. From (4.27) and Lemma 4.5, we obtain

$$\begin{aligned}
\frac{1}{N}\sum_{l=1}^{N}\mathbb{E}\left[\hat{\mathbf{s}}(l)\right]-\boldsymbol{\lambda} &= \frac{1}{N}\sum_{l=1}^{N}\mathbb{E}\left[\mathbf{s}_{\mathbf{v}_l}-\boldsymbol{\lambda}+\hat{\mathbf{s}}(l)-\mathbf{s}_{\mathbf{v}_l}\right] \\
&\geq \frac{1}{N}\sum_{l=1}^{N}\mathbb{E}\left[\mathbf{s}_{\mathbf{v}_l}-\boldsymbol{\lambda}\right]-\rho_2\mathbf{1} \\
&= \frac{1}{N}\sum_{l=1}^{N}\mathbb{E}\left[\frac{\epsilon}{4}\mathbf{1}-\nabla F_\epsilon(\mathbf{v}_l)\right]-\rho_2\mathbf{1}.
\end{aligned}$$

Substituting (B.22) and proceeding, we obtain

$$\frac{1}{N}\sum_{l=1}^{N}\mathbb{E}\left[\hat{\mathbf{s}}(l)\right]-\boldsymbol{\lambda}\geq\left(\frac{\epsilon}{4}-\sqrt{5D\rho_2+\alpha c+\frac{7nD^2}{\alpha N}}-\rho_2\right)\mathbf{1}.$$

Now, choose $\rho_2=\frac{\epsilon^2}{5\times 3^5 D}$. Then,

$$\begin{aligned}
\sqrt{5D\rho_2+\alpha c+\frac{7nD^2}{\alpha N}} &= \sqrt{\frac{\epsilon^2}{3^5}+\frac{\epsilon^2}{3^5}+\frac{\epsilon^2}{3^5}} \\
&= \sqrt{\frac{\epsilon^2}{3^4}}=\frac{\epsilon}{9}. \qquad \text{(B.23)}
\end{aligned}$$

It is easy to check $\rho_2+\frac{\epsilon}{9}\leq\frac{\epsilon}{8}$. This completes the proof. □

B.5.1 Proof of Lemma B.1

Let $\hat{\mathbf{v}} = \mathbf{v} + \Delta\mathbf{v}$. Clearly, $\|\hat{\mathbf{v}}\|_\infty \leq D + 1$. In order to prove $G([\hat{\mathbf{v}}]_D) \geq G(\hat{\mathbf{v}})$, it is sufficient to prove the following. Along any dimension $i \in \mathcal{L}$, $G([\hat{\mathbf{v}}]_{D,i}) \geq G(\hat{\mathbf{v}})$, where $[\hat{\mathbf{v}}]_{D,i}$ is defined as: the i-th component of $[\hat{\mathbf{v}}]_{D,i}$ is same as the i-th component of $[\hat{\mathbf{v}}]_D$, and all other components of $[\hat{\mathbf{v}}]_{D,i}$ are same as the corresponding components of $\hat{\mathbf{v}}$. It is sufficient to prove this as we can repeatedly apply $G([\hat{\mathbf{v}}]_{D,i}) \geq G(\hat{\mathbf{v}})$ along all dimensions to obtain $G([\hat{\mathbf{v}}]_D) \geq G(\hat{\mathbf{v}})$.

Consider any $i \in \mathcal{L}$. If $\hat{v}_i \in [-D, D]$, then $G([\hat{\mathbf{v}}]_{D,i}) = G(\hat{\mathbf{v}})$. Therefore, the only non-trivial cases are $\hat{v}_i \in (D, D+1]$ and $\hat{v}_i \in [-(D+1), -D)$. We consider these cases separately, and apply $|\partial F_\epsilon / \partial v_i| \leq \bar{K}$, and $\|\mathbf{v}^*\|_\infty \leq D - \bar{K}$. For $\hat{v}_i \in (D, D+1]$, we have

$$
\begin{aligned}
&G([\hat{\mathbf{v}}]_{D,i}) - G(\hat{\mathbf{v}}) = \\
&F_\epsilon([\hat{\mathbf{v}}]_{D,i}) - F_\epsilon(\hat{\mathbf{v}}) - ((D - v_i^*)^2 - (\hat{v}_i - v_i^*)^2) \\
&\geq -\bar{K}(\hat{v}_i - D) + (\hat{v}_i - D)(\hat{v}_i + D - 2v_i^*) \\
&\geq (\hat{v}_i - D)(-\bar{K} + \hat{v}_i + D - 2v_i^*) \geq 0.
\end{aligned}
$$

The other case follows from similar arguments.

Since $F_\epsilon(\mathbf{v}) \leq 0$, clearly $G(\mathbf{v}) \leq 0$. Next, we obtain a simple lower

bound on $G(\mathbf{v})$ as follows:

$$\begin{aligned} G(\mathbf{v}) &= F_\epsilon(\mathbf{v}) - \|\mathbf{v} - \mathbf{v}^*\|_2^2 \\ &= (\boldsymbol{\lambda} + \frac{\epsilon}{4}\mathbf{1}) \cdot \mathbf{v} - \log\left(\sum_{\tilde{\mathbf{r}} \in \mathcal{R}} \exp(\tilde{\mathbf{r}} \cdot \mathbf{v})\right) - \|\mathbf{v} - \mathbf{v}^*\|_2^2 \\ &\geq -\bar{K}nD - \log\left(\left\lceil 2\bar{K}/\epsilon \right\rceil^n \exp(\bar{K}nD)\right) - n(2D)^2 \\ &= -n\left(2\bar{K}D + \log\left\lceil 2\bar{K}/\epsilon \right\rceil + 4D^2\right) \\ &\geq -7nD^2. \end{aligned}$$

This completes the proof. □

Appendix C

Proofs for Chapter 7

C.1 Proof of Theorem 7.1

Consider complex Gaussian $\mathcal{CN}(0,1)$ distributions for all inputs Q_{jk}.[1] Then, the resulting mutual information with this (not necessarily maximizing) input distribution $I(Q_{jk};X_{jk})$ is an achievable rate. However, this does not result in a computable expression. To obtain a computable lower bound to this rate, we observe that input random variable Q_{jk} and effective noise $Z_{jk}^{'}$ given by (7.7) are uncorrelated base on the following: The input distributions are such that Q_{jk} is clearly independent of Q_{li} for all $(l,i) \neq (j,k)$ and Z_{jk}. Furthermore, Q_{jk} is independent of G_{jk}^{jk}. Therefore,

$$\mathbb{E}\left[\left(G_{jk}^{jk} - \mathbb{E}\left[G_{jk}^{jk}\right]\right)|Q_{jk}|^2\right] = \mathbb{E}\left[\left(G_{jk}^{jk} - \mathbb{E}\left[G_{jk}^{jk}\right]\right)\right]\mathbb{E}\left[|Q_{jk}|^2\right] = 0.$$

The variance of the effective noise is

$$\mathbb{E}\left[|Z_{jk}^{'}|^2\right] = \text{var}\left\{G_{jk}^{jk}\right\} + \sum_{(l,i)\neq(j,k)} \mathbb{E}\left[|G_{li}^{jk}|^2\right] + 1.$$

Now, from [52] (Theorem 1), we know the result that the channel with independent Gaussian noise $\hat{Z}_{jk}$ with same variance given by

$$\hat{X}_{jk} = \mathbb{E}\left[G_{jk}^{jk}\right]Q_{jk} + \hat{Z}_{jk}$$

[1] Upper case symbols are used in this proof to emphasize that these are random variables.

is worse, i.e.,

$$\begin{aligned} I(Q_{jk}; X_{jk}) &\geq I(Q_{jk}; \hat{X}_{jk}) \\ &= h(\hat{X}_{jk}) - h(\hat{X}_{jk}|Q_{jk}) \\ &= h(\hat{X}_{jk}) - h(\hat{Z}_{jk}) \\ &= \log_2 \left(1 + \frac{\left| \mathbb{E}\left[G_{jk}^{jk} \right] \right|^2}{1 + \mathrm{var}\left\{ G_{jk}^{jk} \right\} + \sum_{(l,i)\neq(j,k)} \mathbb{E}\left[\left| G_{li}^{jk} \right|^2 \right]} \right). \end{aligned}$$

This completes the proof. □

C.2 Proof of Theorem 7.4

The proof of (7.15) follows by substituting the results of Lemma 7.2 and Lemma 7.3 in (7.9). Since θ has a scaled (by a factor of $1/\sqrt{2}$) chi distribution with $2M$ degrees of freedom, it is straightforward to see that

$$\mathbb{E}[\theta] = \frac{\Gamma(M + \frac{1}{2})}{\Gamma(M)},$$

$$\mathbb{E}[\theta^2] = M$$

and

$$\mathrm{var}\{\theta\} = M - \mathbb{E}^2[\theta].$$

Using the duplication formula

$$\Gamma(z)\Gamma\left(z + \frac{1}{2}\right) = 2^{(1-2z)}\sqrt{\pi}\Gamma(2z)$$

and Stirling's formula

$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n} n^n e^{-n}} = 1,$$

we obtain

$$\begin{aligned}
&\lim_{M\to\infty} \frac{1}{\sqrt{M}} \frac{\Gamma\left(M+\frac{1}{2}\right)}{\Gamma(M)} \\
&= \lim_{M\to\infty} \sqrt{\frac{\pi}{M}} 2^{(1-2M)} \frac{(2M-1)!}{(M-1)!(M-1)!}, \\
&= \lim_{M\to\infty} \sqrt{\frac{\pi}{M}} 2^{(1-2M)} \frac{\sqrt{2\pi(2M-1)}(2M-1)^{(2M-1)}e^{1-2M}}{2\pi(M-1)(M-1)^{2(M-1)}e^{2(1-M)}}, \\
&= \lim_{M\to\infty} \sqrt{\frac{2M-1}{2M}} \left(1+\frac{1}{2(M-1)}\right)^{2M-1} e^{-1}, \\
&= 1.
\end{aligned}$$

Therefore,

$$\lim_{M\to\infty} \frac{\mathbb{E}^2[\theta]}{M} = 1$$

and

$$\lim_{M\to\infty} \frac{\mathrm{var}\{\theta\}}{M} = 0.$$

This completes the proof of (7.16). □

C.3 Proof of Theorem 7.5

First, we simplify the objective function $J(\mathbf{A}_l, \alpha_l)$ of the problem (7.18) as follows:

$$\begin{aligned}
J(\mathbf{A}_l, \alpha_l) &= \mathbb{E}\left[\left\| \alpha_l \left(\mathbf{F}_{ll}\mathbf{A}_l\mathbf{q}_l + \mathbf{z}_l \right) - \mathbf{q}_l \right\|^2 + \sum_{j\neq l} \left\| \alpha_l \gamma \mathbf{F}_{jl}\mathbf{A}_l\mathbf{q}_l \right\|^2 \middle| \hat{\mathbf{F}}_{jl} \right], \\
&= \mathbb{E}\left[\left\| \left(\alpha_l \mathbf{F}_{ll}\mathbf{A}_l - \mathbf{I}_K \right) \mathbf{q}_l \right\|^2 + \sum_{j\neq l} \left\| \alpha_l \gamma \mathbf{F}_{jl}\mathbf{A}_l\mathbf{q}_l \right\|^2 \middle| \hat{\mathbf{F}}_{jl} \right] + \alpha_l^2 K.
\end{aligned}$$

Now, by switching linear operators (trace and expectation), we further simplify $J(\mathbf{A}_l, \alpha_l)$:

$$\begin{aligned} J(\mathbf{A}_l, \alpha_l) &= \operatorname{tr}\left\{ \mathbb{E}\left[\left(\alpha_l \mathbf{F}_{ll}\mathbf{A}_l - \mathbf{I}_K\right)^\dagger \left(\alpha_l \mathbf{F}_{ll}\mathbf{A}_l - \mathbf{I}_K\right) \right.\right. \\ &\quad \left.\left. + \sum_{j\neq l} \alpha_l^2 \gamma^2 \mathbf{A}_l^\dagger \mathbf{F}_{jl}^\dagger \mathbf{F}_{jl} \mathbf{A}_l \middle| \hat{\mathbf{F}}_{jl} \right] \right\} + \alpha_l^2 K, \\ &= \operatorname{tr}\left\{ \alpha_l^2 \mathbf{A}_l^\dagger \mathbb{E}\left[\mathbf{F}_{ll}^\dagger \mathbf{F}_{ll} \middle| \hat{\mathbf{F}}_{jl} \right] \mathbf{A}_l + \sum_{j\neq l} \alpha_l^2 \gamma^2 \mathbf{A}_l^\dagger \mathbb{E}\left[\mathbf{F}_{jl}^\dagger \mathbf{F}_{jl} \middle| \hat{\mathbf{F}}_{jl} \right] \mathbf{A}_l \right. \\ &\quad \left. -\alpha_l \mathbf{A}_l^\dagger \hat{\mathbf{F}}_{ll}^\dagger - \alpha_l \hat{\mathbf{F}}_{ll}\mathbf{A}_l \right\} + \left(\alpha_l^2 + 1\right) K, \\ &= \operatorname{tr}\left\{ \alpha_l^2 \mathbf{A}_l^\dagger \left(\hat{\mathbf{F}}_{ll}^\dagger \hat{\mathbf{F}}_{ll} + \gamma^2 \sum_{j\neq l} \hat{\mathbf{F}}_{jl}^\dagger \hat{\mathbf{F}}_{jl} + \left(\delta_{ll} + \gamma^2 \sum_{j\neq l} \delta_{jl} \right) \mathbf{I}_M \right) \mathbf{A}_l \right. \\ &\quad \left. -\alpha_l \mathbf{A}_l^\dagger \hat{\mathbf{F}}_{ll}^\dagger - \alpha_l \hat{\mathbf{F}}_{ll}\mathbf{A}_l \right\} + \left(\alpha_l^2 + 1\right) K. \end{aligned}$$

The last step follows from Lemma C.1.

Now, consider the Lagrangian formulation

$$L\left(\mathbf{A}_l, \alpha_l, \lambda\right) = J\left(\mathbf{A}_l, \alpha_l\right) + \lambda \left(\operatorname{tr}\left\{ \mathbf{A}_l^\dagger \mathbf{A}_l \right\} - 1 \right)$$

for the problem (7.18). Let

$$\mathbf{R} = \hat{\mathbf{F}}_{ll}^\dagger \hat{\mathbf{F}}_{ll} + \gamma^2 \sum_{j\neq l} \hat{\mathbf{F}}_{jl}^\dagger \hat{\mathbf{F}}_{jl} + \left(\delta_{ll} + \gamma^2 \sum_{j\neq l} \delta_{jl} + \frac{\lambda}{\alpha_l^2} \right) \mathbf{I}_M,$$

$\mathbf{U} = \alpha_l \mathbf{R}^{\frac{1}{2}} \mathbf{A}_l$ and $\mathbf{V} = \mathbf{R}^{-\frac{1}{2}} \hat{\mathbf{F}}_{ll}^\dagger$. We have

$$L\left(\mathbf{A}_l, \alpha_l, \lambda\right) = \|\mathbf{U} - \mathbf{V}\|^2 - \operatorname{tr}\left\{ \hat{\mathbf{F}}_{ll} \mathbf{R}^{-1} \hat{\mathbf{F}}_{ll}^\dagger \right\} + \left(\alpha_l^2 + 1\right) K - \lambda. \qquad \text{(C.1)}$$

This can be easily verified by expanding the right hand side. It is clear from (C.1) that, for any given α_l and λ, $L(\mathbf{A}_l, \alpha_l, \lambda)$ is minimized if and only if $\mathbf{U} = \mathbf{V}$. Hence, we obtain

$$\mathbf{A}_l^{opt} = \frac{1}{\alpha_l} \mathbf{R}^{-1} \hat{\mathbf{F}}_{ll}^{\dagger}. \tag{C.2}$$

Let $L(\alpha_l, \lambda) = L(\mathbf{A}_l^{opt}, \alpha_l, \lambda)$. Now, we have

$$L(\alpha_l, \lambda) = -\operatorname{tr}\left\{ \hat{\mathbf{F}}_{ll} \mathbf{R}^{-1} \hat{\mathbf{F}}_{ll}^{\dagger} \right\} + \left(\alpha_l^2 + 1\right) K - \lambda. \tag{C.3}$$

Note that

$$\hat{\mathbf{F}}_{ll}^{\dagger} \hat{\mathbf{F}}_{ll} + \gamma^2 \sum_{j \neq l} \hat{\mathbf{F}}_{jl}^{\dagger} \hat{\mathbf{F}}_{jl}$$

can be factorized in the form

$$\mathbf{S}^{\dagger} \operatorname{diag}\{[c_1 \; c_2 \; \cdots \; c_M]\} \mathbf{S},$$

where $\mathbf{S}^{\dagger}\mathbf{S} = \mathbf{I}_M$. Let $\delta = \delta_{ll} + \gamma^2 \sum_{j \neq l} \delta_{jl}$. Therefore,

$$\begin{aligned}
\mathbf{R}^{-1} &= \left(\mathbf{S}^{\dagger} \operatorname{diag}\{[c_1 \; c_2 \; \cdots \; c_M]\} \mathbf{S} + \left(\delta + \frac{\lambda}{\alpha_l^2} \right) \mathbf{I}_M \right)^{-1}, \\
&= \left(\mathbf{S}^{\dagger} \operatorname{diag} \left\{ \left[c_1 + \delta + \frac{\lambda}{\alpha_l^2} \; c_2 + \delta + \frac{\lambda}{\alpha_l^2} \; \cdots \; c_M + \delta + \frac{\lambda}{\alpha_l^2} \right] \right\} \mathbf{S} \right)^{-1}, \\
&= \mathbf{S}^{\dagger} \operatorname{diag} \left\{ \left[\left(c_1 + \delta + \frac{\lambda}{\alpha_l^2} \right)^{-1} \left(c_2 + \delta + \frac{\lambda}{\alpha_l^2} \right)^{-1} \cdots \right. \right. \\
&\qquad \left. \left. \left(c_M + \delta + \frac{\lambda}{\alpha_l^2} \right)^{-1} \right] \right\} \mathbf{S}.
\end{aligned} \tag{C.4}$$

Substituting (C.4) in (C.3), we get

$$L(\alpha_l, \lambda) = -\sum_{m=1}^{M} \frac{d_m}{c_m + \delta + \frac{\lambda}{\alpha_l^2}} + (\alpha_l^2 + 1)K - \lambda, \tag{C.5}$$

where d_m is the (m,m)-th entry of $\mathbf{S}\hat{\mathbf{F}}_{ll}^{\dagger}\hat{\mathbf{F}}_{ll}\mathbf{S}^{\dagger}$. Consider the equations obtained by differentiating (C.5) w.r.t. α_l and λ and equating to zero:

$$\sum_{m=1}^{M} \frac{d_m}{\left(c_m + \delta + \frac{\lambda}{\alpha_l^2}\right)^2} \frac{1}{\alpha_l^2} = 1, \tag{C.6}$$

$$-\sum_{m=1}^{M} \frac{d_m}{\left(c_m + \delta + \frac{\lambda}{\alpha_l^2}\right)^2} \frac{2\lambda}{\alpha_l^3} + 2\alpha_l K = 0. \tag{C.7}$$

Substituting (C.6) in (C.7), we get

$$\frac{\lambda}{\alpha_l^2} = K. \tag{C.8}$$

Combining the results in (C.2), (C.4), and (C.8), we have

$$\mathbf{A}_l^{opt} = \frac{1}{\alpha_l^{opt}} \left(\hat{\mathbf{F}}_{ll}^{\dagger}\hat{\mathbf{F}}_{ll} + \gamma^2 \sum_{j \neq l} \hat{\mathbf{F}}_{jl}^{\dagger}\hat{\mathbf{F}}_{jl} + \left(\delta_{ll} + \gamma^2 \sum_{j \neq l} \delta_{jl} + K \right) \mathbf{I}_M \right)^{-1} \hat{\mathbf{F}}_{ll}^{\dagger},$$

where α_l^{opt} is such that $\|\mathbf{A}_l^{opt}\|^2 = 1$. This completes the proof. □

Lemma C.1. *Consider the optimization problem (7.18). For all j and l,*

$$\mathbb{E}\left[\tilde{\mathbf{F}}_{jl}^{\dagger}\tilde{\mathbf{F}}_{jl} \middle| \hat{\mathbf{F}}_{jl} \right] = \delta_{jl}\mathbf{I}_M,$$

where

$$\delta_{jl} = p_f \operatorname{tr}\left\{ \mathbf{D}_{jl} \left(\mathbf{I}_K + p_r \tau \mathbf{D}_{jl}^{\frac{1}{2}} \mathbf{\Psi}_j^{\dagger} \mathbf{\Lambda}_{jl} \mathbf{\Psi}_j \mathbf{D}_{jl}^{\frac{1}{2}} \right)^{-1} \right\}, \tag{C.9}$$

and

$$\mathbf{\Lambda}_{jl} = \left(\mathbf{I} + p_r \tau \sum_{i \neq j} \mathbf{\Psi}_i \mathbf{D}_{il} \mathbf{\Psi}_i^{\dagger} \right)^{-1}.$$

Proof. Let $\tilde{\mathbf{f}}_{jlm}$ denote the m-th column of $\tilde{\mathbf{F}}_{jl}$. Similarly, we define $\mathbf{h}_{jlm}$ and $\hat{\mathbf{h}}_{jlm}$. From (7.3), we have

$$\begin{aligned}\tilde{\mathbf{f}}_{jlm} &= \sqrt{p_f}\mathbf{D}_{jl}^{\frac{1}{2}}(\mathbf{h}_{jlm} - \hat{\mathbf{h}}_{jlm}), \\ &= \sqrt{p_f}\mathbf{D}_{jl}^{\frac{1}{2}}\left(\mathbf{h}_{jlm} - \sqrt{p_r\tau}\mathbf{D}_{jl}^{\frac{1}{2}}\mathbf{\Psi}_j^{\dagger}\left(\mathbf{I} + p_r\tau\sum_{i=1}^{L}\mathbf{\Psi}_i\mathbf{D}_{il}\mathbf{\Psi}_i^{\dagger}\right)^{-1}\mathbf{y}_{lm}\right),\end{aligned}$$

where $\mathbf{y}_{lm}$ is given by (7.1). For given j and l, it is clear that $\left\{\tilde{\mathbf{f}}_{jlm}\right\}_{m=1}^{M}$ is i.i.d. zero-mean $\mathcal{CN}$ distributed. Hence,

$$\mathbb{E}\left[\tilde{\mathbf{F}}_{jl}^{\dagger}\tilde{\mathbf{F}}_{jl}\middle|\hat{\mathbf{F}}_{jl}\right] = \delta_{jl}\mathbf{I}_M$$

with

$$\begin{aligned}\delta_{jl} &= \mathbb{E}\left[\tilde{\mathbf{f}}_{jlm}^{\dagger}\tilde{\mathbf{f}}_{jlm}\middle|\hat{\mathbf{F}}_{jl}\right], \\ &= p_f\,\mathrm{tr}\left\{\mathbf{D}_{jl}^{\frac{1}{2}}\left(\mathbf{I}_K - \mathbb{E}\left[\hat{\mathbf{h}}_{jlm}\hat{\mathbf{h}}_{jlm}^{\dagger}\middle|\hat{\mathbf{F}}_{jl}\right]\right)\mathbf{D}_{jl}^{\frac{1}{2}}\right\}, \\ &= p_f\,\mathrm{tr}\left\{\mathbf{D}_{jl}^{\frac{1}{2}}\left(\mathbf{I}_K - p_r\tau\mathbf{D}_{jl}^{\frac{1}{2}}\mathbf{\Psi}_j^{\dagger}\mathbf{\Lambda}_{jl}\mathbf{\Psi}_j\mathbf{D}_{jl}^{\frac{1}{2}}\right)\mathbf{D}_{jl}^{\frac{1}{2}}\right\}, \\ &= p_f\,\mathrm{tr}\left\{\mathbf{D}_{jl}^{\frac{1}{2}}\left(\mathbf{I}_K + p_r\tau\mathbf{D}_{jl}^{\frac{1}{2}}\mathbf{\Psi}_j^{\dagger}\mathbf{\Lambda}_{jl}\mathbf{\Psi}_j\mathbf{D}_{jl}^{\frac{1}{2}}\right)^{-1}\mathbf{D}_{jl}^{\frac{1}{2}}\right\},\end{aligned}$$

where

$$\mathbf{\Lambda}_{jl} = \left(\mathbf{I} + p_r\tau\sum_{i\neq j}\mathbf{\Psi}_i\mathbf{D}_{il}\mathbf{\Psi}_i^{\dagger}\right)^{-1}.$$

The last step follows from matrix inversion lemma. This completes the proof of the lemma. □

Bibliography

[1] 3GPP Long Term Evolution Coordinated Multi-Point transmission and reception (LTE CoMP). http://www.3gpp.org/.

[2] Cisco Visual Networking Index: Forecast and Methodology, 2009-2014. http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360.pdf.

[3] Federal Communications Commission. http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-08-260A1.pdf.

[4] R. Ahlswede. Multi-way communication channels. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, pages 103–135, 1971.

[5] M. Airy, S. Bhadra, R. W. Heath, Jr., and S. Shakkottai. Transmit precoding for the multiple antenna broadcast channel. In *Proc. IEEE Vehicular Technology Conference (VTC)*, volume 3, pages 1396–1400, 2006.

[6] J. Andrews, A. Ghosh, and R. Muhamed. *Fundamentals of WiMAX*. Prentice Hall USA, 2007.

[7] V. Annapureddy and V. Veeravalli. Gaussian interference networks: Sum capacity in the low-interference regime and new outer bounds on

the capacity region. *IEEE Trans. Inform. Theory*, 55(7):3032–3050, 2009.

[8] A. Ashikhmin and R. Gopalan. Grassmannian packings for efficient quantization in MIMO broadcast systems. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, pages 1811 –1815, Jun. 2007.

[9] S. Asmussen. *Applied Probability and Queues*. Springer-Verlag, New York, 2003.

[10] T. Berger. *Rate distortion theory*. Prentice-Hall Englewood Cliffs, NJ, 1971.

[11] V. Bhargava, J. Jose, K. Srinivasan, and S. Vishwanath. Q-CMRA: Queue-based channel measurement & rate allocation. Technical Report TR-ARiSE-2011-004, UT Austin, Feb. 2011.

[12] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and H. V. Poor. *MIMO Wireless Communications*. Cambridge University Press, 2010.

[13] F. Boccardi, F. Tosato, and G. Caire. Precoding schemes for the MIMO-GBC. In *Proc. IEEE International Zurich Seminar on Communications*, Feb. 2006.

[14] C. Bordenave, D. McDonald, and A. Proutiere. Performance of random medium access control, an asymptotic approach. In *Proc. ACM Sigmetrics*, 2008.

[15] J. Borwein and A. Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer Verlag, 2006.

[16] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

[17] P. Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. Springer, 1999.

[18] V. R. Cadambe and S. A. Jafar. Interference alignment and the degrees of freedom for the K-user interference channel. *IEEE Trans. Inform. Theory*, 54(8):3425–3441, Aug. 2008.

[19] V. R. Cadambe and S. A. Jafar. Multiple access outerbounds and the inseparability of parallel Gaussian interference channels. In *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, 2008.

[20] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran. Multiuser MIMO achievable rates with downlink training and channel state feedback. *IEEE Trans. Inform. Theory*, 56(6):2845–2866, Jun. 2010.

[21] G. Caire and S. Shamai (Shitz). On the achievable throughput of a multiantenna Gaussian broadcast channel. *IEEE Trans. Inform. Theory*, 49:1691–1707, Jul. 2003.

[22] A. B. Carleial. A case where interference does not reduce capacity. *IEEE Trans. Inform. Theory*, 21:569, 1975.

[23] J. Chakareski and P. Frossard. Rate-distortion optimized distributed packet scheduling of multiple video streams over shared communication resources. *IEEE Trans. Multimedia*, 8(2):207, 2006.

[24] D. Chen, K. Azarian, and J. N. Laneman. A case for amplify-forward relaying in the block-fading multiple-access channel. *IEEE Trans. Inform. Theory*, 54(8):3728–3733, 2008.

[25] M. Chiang. Balancing transport and physical layers in wireless multihop networks: Jointly optimal congestion control and power control. *IEEE J. Sel. Areas Commun.*, 23(1), 2005.

[26] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle. Layering as optimization decomposition: A mathematical theory of network architectures. *Proceedings of the IEEE*, 95(1):255–312, March 2007.

[27] P. Chou and Z. Miao. Rate-distortion optimized streaming of packetized media. *IEEE Trans. Multimedia*, 8(2), 2006.

[28] S. Christensen, R. Agarwal, E. Carvalho, and J. Cioffi. Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design. *IEEE Trans. Wireless Commun.*, 7(12):1–8, 2008.

[29] M. Costa and A. El-Gamal. The capacity region of the discrete memoryless interference channel with strong interference. *IEEE Trans. Inform. Theory*, 33(5):710–711, 1987.

[30] T. Cover and J. Thomas. *Elements of information theory.* Wiley-Interscience, 2006.

[31] T. M. Cover and A. El-Gamal. Capacity theorems for the relay channel. *IEEE Trans. Inform. Theory*, 25(5):572–584, 1979.

[32] T. Cui, T. Ho, and L. Chen. On distributed distortion optimization for correlated sources. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, 2008.

[33] J. Dai. On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probab.*, 5(1):49–77, 1995.

[34] S. Deb, V. Srinivasan, and R. Maheshwari. Dynamic spectrum access in DTV whitespaces: Design rules, architecture and algorithms. In *Proc. ACM MobiCom*, 2009.

[35] P. Ding, D. J. Love, and M. D. Zoltowski. On the sum rate of channel subspace feedback for multi-antenna broadcast channels. In *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, pages 2699–2703, Nov. 2005.

[36] L. Dong. *Cross-layer design for cooperative wireless networks.* PhD thesis, Drexel University, 2008.

[37] Y. Eldar and N. Merhav. A competitive minimax approach to robust estimation of random parameters. *IEEE Trans. Signal Process.*, 52(7):1931, 2004.

[38] A. Ephremides and B. Hajek. Information theory and communication networks: An unconsummated union. *IEEE Trans. Inform. Theory*, 44(6):2416–2434, 1998.

[39] A. Eryilmaz, R. Srikant, and J. R. Perkins. Stable scheduling policies for fading wireless channels. *IEEE/ACM Trans. Netw.*, 13(2):411–424, 2005.

[40] R. Etkin and E. Ordentlich. On the degrees-of-freedom of the K-user Gaussian interference channel. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, pages 1919–1923, 2009.

[41] R. Etkin, D. Tse, and H. Wang. Gaussian interference channel capacity to within one bit. *IEEE Trans. Inform. Theory*, 54(12):5534–5562, 2008.

[42] F. H. Fitzek and M. D. Katz. *Cognitive Wireless Networks*. Springer, 2007.

[43] G. J. Foschini, K. Karakayali, and R. A. Valenzuela. Coordinating multiple antenna cellular networks to achieve enormous spectral efficiency. *IEE Proceedings Communications*, 153:548–555, Aug. 2006.

[44] L. Georgiadis, M. J. Neely, and L. Tassiulas. *Resource Allocation and Cross-Layer Control in Wireless Networks*. Now Publishers, 2006.

[45] K. Gomadam, V. Cadambe, and S. Jafar. Approaching the capacity of wireless networks through distributed interference alignment. In *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, 2008.

[46] K. S. Gomadam, H. C. Papadopoulos, and C.-E. W. Sundberg. Techniques for multi-user MIMO with two-way training. In *Proc. IEEE International Conference on Communications (ICC)*, pages 3360–3366, May 2008.

[47] T. Gou and S. Jafar. Degrees of freedom of the K user M times N MIMO interference channel. *IEEE Trans. Inform. Theory*, 56(12):6040–6057, 2010.

[48] V. Goyal. Multiple description coding: Compression meets the network. *IEEE Signal Processing Magazine*, 18(5):74–93, 2001.

[49] M. Guillaud, D. T. M. Slock, and R. Knopp. A practical method for wireless channel reciprocity exploitation through relative calibration. In *Proc. ISSPA*, 2005.

[50] P. Gupta and A. L. Stolyar. Optimal throughput allocation in general random-access networks. In *Proc. Info. Sciences and Systems (CISS)*, pages 1254–1259, 2006.

[51] T. Han and K. Kobayashi. A new achievable rate region for the interference channel. *IEEE Trans. Inform. Theory*, 27(1):49–60, 1981.

[52] B. Hassibi and B. M. Hochwald. How much training is needed in multiple-antenna wireless links? *IEEE Trans. Inform. Theory*, 49:951–963, Apr. 2003.

[53] B. M. Hochwald, C. B. Peel, and A. L. Swindlehurst. A vector-perturbation technique for near-capacity multiantenna multiuser communication part II: Perturbation. *IEEE Trans. Commun.*, 53:537–544, Jan. 2005.

[54] K. Huang, R. W. Heath, Jr., and J. G. Andrews. Space division multiple access with a sum feedback rate constraint. *IEEE Trans. Signal Process.*, 55(7):3879–3891, 2007.

[55] S. A. Jafar. The ergodic capacity of interference networks. `http://arxiv.org/abs/0902.0838`, 2009.

[56] A. Jafarian, J. Jose, and S. Vishwanath. Algebraic lattice alignment for K-user interference channels. In *Proc. Allerton Conference on Commun., Control and Computing*, pages 88–93, Oct. 2009.

[57] L. Jiang, D. Shah, J. Shin, and J. Walrand. Distributed random access algorithm: Scheduling and congestion control. *IEEE Trans. Inform. Theory*, 56(12):6182–6207, 2010.

[58] L. Jiang and J. Walrand. Convergence analysis of a distributed CSMA algorithm for maximal throughput in a general class of networks. Technical Report UCB/EECS-2008-185, UC Berkeley, Dec. 2008.

[59] L. Jiang and J. Walrand. A distributed CSMA algorithm for throughput and utility maximization in wireless networks. In *Proc. Allerton Conference on Commun., Control and Computing*, 2008.

[60] L. Jiang and J. Walrand. Approaching throughput-optimality in a distributed CSMA algorithm with contention resolution. Technical Report UCB/EECS-2009-37, UC Berkeley, Mar. 2009.

[61] N. Jindal. MIMO broadcast channels with finite-rate feedback. *IEEE Trans. Inform. Theory*, 52:5045–5060, Nov. 2006.

[62] M. Joham, P. M. Castro, L. Castedo, and W. Utschick. MMSE optimal feedback of correlated CSI for multi-user precoding. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3129–3132, Mar. 2008.

[63] J. Jose, A. Ashikhmin, T. Marzetta, and S. Vishwanath. Pilot contamination problem in multi-cell TDD systems. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, pages 2184 –2188, Jul. 2009.

[64] J. Jose, A. Ashikhmin, T. Marzetta, and S. Vishwanath. Pilot contamination and precoding in multi-cell TDD systems. *Accepted for publication in IEEE Trans. on Wireless Commun.*, 2011.

[65] J. Jose, A. Ashikhmin, P. Whiting, and S. Vishwanath. Scheduling and pre-conditioning in multi-user MIMO TDD systems. In *Proc. IEEE International Conference on Communications (ICC)*, pages 4100 – 4105, May 2008.

[66] J. Jose, A. Ashikhmin, P. Whiting, and S. Vishwanath. Channel estimation and linear precoding in multiuser multiple-antenna TDD systems. *IEEE Trans. on Veh. Technol.*, 60(5):2102–2116, Jun. 2011.

[67] J. Jose, N. Prasad, M. Khojastepour, and S. Rangarajan. On robust weighted-sum rate maximization in MIMO interference networks. In *Proc. IEEE International Conference on Communications (ICC)*, Jun. 2011.

[68] J. Jose and S. Vishwanath. Sum capacity of K user Gaussian degraded interference channels. In *Proc. IEEE Info. Theory Workshop (ITW)*, Dublin, Ireland, Sep. 2010.

[69] J. Jose and S. Vishwanath. Distributed rate allocation for wireless networks. *Accepted for publication in IEEE Trans. Inform. Theory*, 2011.

[70] J. Jose and S. Vishwanath. Network control: A rate-distortion perspective. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, Aug. 2011.

[71] J. Jose, L. Ying, and S. Vishwanath. On the stability region of amplify-and-forward cooperative relay networks. In *Proc. IEEE Info. Theory Workshop (ITW)*, pages 620–624, Taormina, Sicily, Oct. 2009.

[72] T. Kailath, A. H. Sayed, and B. Hassibi. *Linear Estimation*. Prentice Hall, 2000.

[73] M. Kalman, P. Ramanathan, and B. Girod. Rate-distortion optimized video streaming with multiple deadlines. In *Proc. of International Conference on Image Processing (ICIP)*, 2003.

[74] F. Kelly. Mathematical modelling of the internet. *Mathematics Unlimited-2001 and Beyond*, pages 685–702, 2001.

[75] M. Khojastepour, X. Wang, and M. Madihian. Design of multiuser downlink linear MIMO precoding systems with quantized feedback. *IEEE Trans. on Veh. Technol.*, 58(9):4828 –4836, Nov. 2009.

[76] G. Kramer. Outer bounds on the capacity of Gaussian interference channels. *IEEE Trans. Inform. Theory*, 50:581–586, Mar. 2004.

[77] J. Laneman, D. Tse, and G. Wornell. Cooperative diversity in wireless networks: Efficient protocols and outage behavior. *IEEE Trans. Inform. Theory*, 50(12):3062–3080, 2004.

[78] J. Lee, J. Lee, Y. Yi, S. Chong, A. Proutiere, and M. Chiang. Implementing utility optimal CSMA. In *Proc. Allerton Conference on Commun., Control and Computing*, 2009.

[79] Y. Li. *Optimal network resource allocation for heterogeneous traffic.* PhD thesis, Princeton University, Princeton, NJ, USA, 2008.

[80] X. Lin, N. B. Shroff, and R. Srikant. A tutorial on cross-layer optimization in wireless networks. *IEEE J. Sel. Areas Commun.*, 24(8):1452–1463, July 2006.

[81] J. Liu, Y. Yi, A. Proutiere, M. Chiang, and H. V. Poor. Convergence and tradeoff of utility-optimal CSMA. *Wiley Journal of Wireless Communications and Mobile Computing, Special Issue on Advances in Wireless Communications and Networking*, Dec. 2009.

[82] Z. Luo and S. Zhang. Dynamic spectrum management: Complexity and duality. *IEEE J. Sel. Topics Signal Process.*, 2(1):57–73, 2008.

[83] P. Marbach, A. Eryilmaz, and A. Ozdaglar. Achievable rate region of CSMA schedulers in wireless networks with primary interference constraints. In *Proc. IEEE Conf. on Decision and Control*, 2007.

[84] T. L. Marzetta. How much training is required for multiuser MIMO? In *Proc. Asilomar Conference on Signals, Systems and Computers (ACSSC)*, pages 359–363, Nov. 2006.

[85] T. L. Marzetta and B. M. Hochwald. Fast transfer of channel state information in wireless systems. *IEEE Trans. Signal Process.*, 54:1268–1278, 2006.

[86] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.

[87] J. Mitola and G. Q. Maguire. Cognitive radio: Making software radios more personal. *IEEE Personal Communications*, 6(4):13–18, 1999.

[88] A. Motahari and A. Khandani. Capacity bounds for the Gaussian interference channel. *IEEE Trans. Inform. Theory*, 55(2):620 –643, Feb. 2009.

[89] A. S. Motahari, S. O. Gharan, M. A. Maddah-Ali, and A. K. Khandani. Real interference real interference alignment: Exploring the potential of single antenna systems. `http://arxiv.org/abs/0908.2282/`, 2009.

[90] B. Nazer, M. Gastpar, S. Jafar, and S. Vishwanath. Ergodic interference alignment. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, pages 1769–1773, 2009.

[91] J. Ni, B. Tan, and R. Srikant. Q-CSMA: Queue-length based CSMA/CA algorithms for achieving maximum throughput and low delay in wireless networks. In *Proc. IEEE Infocom*, Mar. 2010.

[92] A. Nosratinia, T. Hunter, and A. Hedayat. Cooperative communication in wireless networks. *IEEE Communications Magazine*, 42(10):74–80, 2004.

[93] D. O'Neill, E. Akuiyibo, S. Boyd, and A. Goldsmith. Optimizing adaptive modulation in wireless networks via multi-period network utility

maximization. In *Proc. IEEE International Conference on Communications (ICC)*, 2010.

[94] D. O'Neill, A. Goldsmith, and S. Boyd. Cross-layer design with adaptive modulation: Delay, rate, and energy tradeoffs. In *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, 2008.

[95] Y. Oohama. The rate-distortion function for the quadratic Gaussian CEO problem. *IEEE Trans. Inform. Theory*, 44(3):1057, 1998.

[96] D. Palomar and M. Chiang. A tutorial on decomposition methods for network utility maximization. *IEEE J. Sel. Areas Commun.*, 24(8):1439–1451, 2006.

[97] D. Papailiopoulos and A. Dimakis. Interference alignment as a rank constrained rank minimization. In *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, Dec. 2010.

[98] S. Peters and R. W. Heath, Jr. Cooperative algorithms for mimo interference channels. *IEEE Trans. on Veh. Technol.*, 60(1):206–218, Jan. 2011.

[99] S. Rajagopalan, D. Shah, and J. Shin. Network adiabatic theorem: an efficient randomized protocol for contention resolution. In *Proc. ACM Sigmetrics*, pages 133–144, 2009.

[100] I. Santamaria, O. Gonzalez, R. W. Heath, Jr., and S. Peters. Maximum sum-rate interference alignment algorithms for MIMO channels.

In *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, Dec. 2010.

[101] I. Sason. On achievable rate regions for the Gaussian interference channel. *IEEE Trans. Inform. Theory*, 50(6):1345 – 1356, Jun. 2004.

[102] H. Sato. On degraded gaussian two-user channels (corresp.). *IEEE Trans. Inform. Theory*, 24(5):637 – 640, sep 1978.

[103] H. Sato. The capacity of the Gaussian interference channel under strong interference (corresp.). *IEEE Trans. Inform. Theory*, 27:786–788, 1981.

[104] B. Schein. *Distributed Coordination in Network Information Theory*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 2001.

[105] D. Schmidt, C. Shi, R. Berry, M. Honig, and W. Utschick. Minimum mean squared error interference alignment. In *Proc. of Asilomar Conference on Signals, Systems, and Computers*, 2009.

[106] H. Seferoglu, A. Markopoulou, and U. Kozat. Network coding-aware rate control and scheduling in wireless networks. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1496–1499, 2009.

[107] A. Sendonaris, E. Erkip, and B. Aazhang. User cooperation diversity part I and part II. *IEEE Trans. Commun.*, 51(11):1927–48, Nov. 2003.

[108] E. Setton, T. Yoo, X. Zhu, A. Goldsmith, and B. Girod. Cross-layer design of ad hoc networks for real-time video streaming. *IEEE Wireless Communications*, 12(4):59–65, 2005.

[109] D. Shah, D. Tse, and J. N. Tsitsiklis. Hardness of low delay network scheduling. In *IEEE Info. Theory Workshop (ITW)*, Jan. 2010.

[110] S. Shakkottai and R. Srikant. Network optimization and control. *Foundations and Trends® in Networking*, 2(3), 2007.

[111] S. Shamai (Shitz), O. Somekh, and B. M. Zaidel. Multi-cell communications: An information-theoretic perspective. In *Proc. Joint Workshop on Communicaitons and Coding (JWCC)*, Florence, Italy, Oct. 2004.

[112] X. Shang, G. Kramer, and B. Chen. A new outer bound and the noisy-interference sum-rate capacity for Gaussian interference channels. *IEEE Trans. Inform. Theory*, 55(2):689–699, 2009.

[113] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(7):379–423, 1948.

[114] M. Sharif and B. Hassibi. On the capacity of MIMO broadcast channels with partial side information. *IEEE Trans. Inform. Theory*, 51:506–522, Feb. 2005.

[115] Z. Shen, R. Chen, J. G. Andrews, R. W. Heath, Jr., and B. L. Evans. Low complexity user selection algorithms for multiuser MIMO systems

with block diagonalization. *IEEE Trans. Signal Process.*, 54:3658–3663, Sep. 2006.

[116] M. Stojnic, H. Vikalo, and B. Hassibi. Rate maximization in multi-antenna broadcast channels with linear preprocessing. *IEEE Trans. Wireless Commun.*, 5:2338–2342, Sep. 2006.

[117] A. L. Stolyar. Dynamic distributed scheduling in random access networks. *J. Appl. Probab.*, 45(2):297–313, 2008.

[118] A. Tajer, N. Prasad, and X. Wang. Robust beamforming for multi-cell downlink transmission. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, Jun. 2010.

[119] L. Tassiulas and A. Ephremides. Jointly optimal routing and scheduling in packet radio networks. *IEEE Trans. Inform. Theory*, 38(1):165, 1992.

[120] L. Tassiulas and A. Ephremides. Stability properties of constrained queuing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Trans. Autom. Control*, 37(12):1936–1948, Dec. 1992.

[121] C. Tian, J. Chen, S. Diggavi, and S. Shamai. Optimality and approximate optimality of source-channel separation in networks. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, Jun. 2010.

[122] L. Vandenberghe, S. Boyd, and S. Wu. Determinant maximization with linear matrix inequality constraints. *SIAM journal on matrix analysis and applications*, 19:499–533, 1998.

[123] S. Venkatesan, A. Lozano, and R. Valenzuela. Network MIMO: Overcoming intercell interference in indoor wireless systems. In *Proc. Asilomar Conference on Signals, Systems and Computers (ACSSC)*, pages 83–87, Nov. 2007.

[124] S. Vishwanath, N. Jindal, and A. J. Goldsmith. Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels. *IEEE Trans. Inform. Theory*, 49:2658–2668, Oct. 2003.

[125] P. Viswanath and D. N. C. Tse. Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality. *IEEE Trans. Inform. Theory*, 49:1912–1921, Aug. 2003.

[126] B. Vojčić and W. M. Jang. Transmitter precoding in synchronous multiuser communications. *IEEE Trans. Commun.*, 46(10):1346 –1355, Oct. 1998.

[127] N. Vučic, H. Boche, and S. Shi. Robust transceiver optimization in downlink multiuser MIMO systems. *IEEE Trans. Signal Process.*, 57(9):3576–3587, 2009.

[128] A. Wagner, S. Tavildar, and P. Viswanath. The rate region of the quadratic Gaussian two-terminal source-coding problem. *IEEE Trans.*

Inform. Theory, 54(5):1938–1961, 2008.

[129] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz). The capacity region of the Gaussian multiple-input multiple-output broadcast channel. *IEEE Trans. Inform. Theory*, 52:3936–3964, Sep. 2006.

[130] Y. Xi and E. Yeh. Optimal capacity allocation, routing, and congestion control in wireless networks. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, pages 2511–2515, 2006.

[131] Y. Xi and E. Yeh. Distributed algorithms for spectrum allocation, power control, routing, and congestion control in wireless networks. In *Proc. ACM MobiHoc*, pages 180–189, 2007.

[132] E. Yeh and R. Berry. Throughput optimal control of cooperative relaying networks. *IEEE Trans. Inform. Theory*, 53:3827–3832, Oct. 2007.

[133] E. Yeh and R. Berry. Throughput optimal control of wireless networks with two-hop cooperative relaying. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, Jun. 2007.

[134] E. M. Yeh and A. S. Cohen. Delay optimal rate allocation in multiaccess fading communications. In *Proc. of Allerton Conference on Communication, Control, and Computing*, Oct. 2004.

[135] L. Ying, R. Srikant, and D. Towsley. Cluster-based back-pressure routing algorithm. In *Proc. IEEE Infocom*, pages 484–492, 2008.

[136] T. Yoo, N. Jindal, and A. Goldsmith. Multi-antenna broadcast channels with limited feedback and user selection. *IEEE J. Sel. Areas Commun.*, 25:1478–1491, 2007.

[137] W. Yu and J. M. Cioffi. Sum capacity of Gaussian vector broadcast channels. *IEEE Trans. Inform. Theory*, 50(9):1875–1892, 2004.

[138] W. Yu, W. Rhee, S. Boyd, and J. Cioffi. Iterative water-filling for Gaussian vector multiple-access channels. *IEEE Trans. Inform. Theory*, 50(1):145, 2004.

[139] H. Zhang and H. Dai. Co-channel interference mitigation and cooperative processing in downlink multicell multiuser MIMO networks. *European Journal on Wireless Communications and Networking*, pages 222–235, 4th Quarter 2004.

[140] J. Zhang, R. Chen, J. G. Andrews, A. Ghosh, and R. W. Heath, Jr. Networked MIMO with clustered linear precoding. *Trans. Wireless. Comm.*, 8(4):1910–1921, 2009.

Vita

Jubin Jose is a doctoral candidate in the department of Electrical and Computer Engineering at The University of Texas at Austin. He received his B.Tech. in Electrical Engineering from the Indian Institute of Technology (IIT) Madras in 2006 and M.S. in Electrical and Computer Engineering from UT Austin in 2008. His industry experience includes working as Research Assistant at NEC labs, Princeton, NJ; Interim Engineering Intern at Qualcomm Corporate R&D, San Diego, CA; and Consultant at Alcatel-Lucent Bell Labs, Murray Hill, NJ. His research interests lie in the broad area of communication networks.

Permanent address: jubinj@gmail.com

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.