Resource Allocation and Design Issues in Wireless Systems

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Rohit Aggarwal, B. Tech, M. S.

Graduate Program in Department of Electrical and Computer Engineering

The Ohio State University

2012

Dissertation Committee:

Phil Schniter, Advisor

Can Emre Koksal, Co-Advisor

Ness B. Shroff

Abstract

Present wireless communication systems are required to support a variety of high-speed data communication services for its users, such as video streaming and cloud-based services. As the users' demands for such services grow, more efficient wireless systems need to be designed that can support high-speed data and, at the same time, serve the users in a fair manner. One method to achieve this is by using efficient resource allocation schemes at the transmitters of wireless communication systems. Here, the term "resources" refers to the fundamental physical and network-layer quantities that limit the amount of data that can be transmitted over a communication link, such as available bandwidth and power. With the above fact in mind, we study, in this dissertation, physical and network-layer resource allocation problems in three different communication systems, and propose various optimal and sub-optimal solutions. First, we consider point-to-point links and propose greedy low and high-complexity rate-assignment schemes using degraded feedback (ACK/NAK) to maximize goodput of the system. Here, goodput is defined as the amount of data that a transmitter can send to the receiver without any error. Second, we propose optimal and sub-optimal schemes for simultaneous user-scheduling, power-allocation, and rate-selection in an Orthogonal Frequency Division Multiple Access (OFDMA) downlink, with the goal of maximizing expected sum-utility under a sum-power constraint and the availability of imperfect channel-state information at the

transmitter. Finally, we consider allocation of network resources in the downlink of large multi-cellular OFDMA-networks and answer, from the view-point of a service provider, complex questions related to network design such as: How many base-stations need to be installed, or how much bandwidth should be purchased given a particular revenue-target, or what pricing scheme must be employed to meet the Quality-of-Service constraints of users and, at the same time, maximize revenue for the service provider? Using our analysis, we also propose a near-optimal resource allocation scheme for multi-cellular OFDMA systems under a truncated path-loss model and different fading models (Rayleigh, Nakagami-m, Weibull, and LogNormal) that is optimal in the scaling sense.

Dedicated to *Buaji* and my family.

Acknowledgments

I had a wonderful time both inside and outside my workplace, Information Processing Systems (IPS) lab, at The Ohio State University (OSU). OSU is truly an amazing university and I am grateful to God for this unforgettable experience.

First and foremost, I thank my advisors Prof. Philip Schniter and Prof. Can Emre Koksal for their invaluable guidance, support, and belief in me that made this dissertation possible. Phil and Emre inspired me by example and showed how to be creative and perfectionist in work. Their creativity and fundamental thinking made complex ideas so simple and invariably gave rise to new interesting ideas. Often, I used to come out of our meetings thinking "Why did I not think of that?". I am grateful to them for providing me the opportunity of research and encouraging me to pursue my own ideas. Many thanks go to Prof. Ness Shroff for being in my candidacy and dissertation committees and for his constructive comments. I also thank Phil and Prof. Lee Potter for giving me an opportunity to develop and teach the communication lab course *ECE 508* for undergraduate students.

I want to express my gratitude to my *Buaji* and my family - *Papaji, Mummy, Tauji, Taiji, Dadiji, Bhaiya, Mahak, and Chintu* - for their unconditional love, support, and inspiration to constantly strive for improvement in every aspect and in all stages of life. Moments spend with them will always be remembered and cherished by me.

During my doctoral education, I saw many friends come to OSU and leave with different degrees. I want to thank all my friends at OSU with whom I spent time and had lots of fun during the past 4 years. Special thanks go to my best friends and roommates - Anupam Vivek, Suryarghya Chakrabarti, and Vinod Khare - for not complaining (too much) about the food cooked by the worst chef of the house (me) and making *237 West 9th Ave* an all-time "party" house. I also thank *La Familia*, members of which were my friends living near my house, for an amazing time spent together in parties, fun-trips, and games of cricket, soccer, tennis and poker.

I thank my labmates - Justin "the boy with the cup" Ziniel, Rahul Srivastava, Arun Sridharan, Ahmed Fasih, and Mohammad Shahmohammadi for our intense discussions, homework solving sessions, lunch and happy-hour breaks, and comedy in IPS lab. Discussions with them were always helpful in the furthering research whenever it seemed to get stuck. Thanks also to Sibasish Das, Subhojit Som, Liza Toher, Sung-Jun Hwang, Young-Han Nam, and Sugumar Murugesan for our discussions and casual conversations. Their friendliness and spontaneity made the lab a fun-place to work. I thank Carl Rossler, Justin, Rahul, Arun, Brian Caroll, Bo Ji, Harsha Gangammanavar, Subhash Lakshminarayana, and Wenzhuo Owen Ouyang for the homework solving, doubt-clearing sessions, and (sometimes) tennis breaks. I also thank the newbies, Brian Day and Jeremy Vila, for being my students in the undergraduate class *ECE 508* and now colleagues in IPS lab. It is extremely satisfying to teach talented students and see them pursue the path of research. I thank Jeri McMichael for her friendliness, outstanding organizational skills to get the nerds of IPS lab participate in fun-picnics and lab-cleaning exercises, and bearing up with my

mistakes in financial activities such as ordering items for labs, conference registrations, travel reservations, and reimbursements.

Finally, I want to thank Ankita "AK47" Agarwal for her warmth, encouragement, and emotional support. Her charming personality and her cooking always made my day!

Vita

March 15, 1986 ..... Born - Ranchi, Jharkhand, India

2007 ..... B.Tech. Indian Institute of Technology Kanpur

2011 ..... M.S. The Ohio State University

Publications

Research Publications

R. Aggarwal, C. E. Koksal, and P. Schniter "On the Design of Large Scale Wireless Systems". submitted, *IEEE Journal on Selected Areas on Communications - Large Scale Multiple Antenna Wireless Systems*, February 2012

R. Aggarwal, C. E. Koksal, and P. Schniter "Joint Scheduling and Resource Allocation in OFDMA Downlink Systems via ACK/NAK Feedback", *IEEE Transactions on Signal Processing*, Volume 60, Issue 6, pp. 3217-3227, June 2012.

R. Aggarwal, M. Assaad, C. E. Koksal, and P. Schniter "Joint Scheduling and Resource Allocation in OFDMA Downlink Systems with Imperfect Channel-State Information," *IEEE Transactions on Signal Processing*, Volume 59, Issue 11, pp. 5589-5604, November 2011.

R. Aggarwal, P. Schniter, and C. E. Koksal "Rate Adaptation via link-layer feedback for goodput maximization over a Time-Varying Channels," *IEEE Transactions on Wireless Communications*, Volume 8, Issue 8, pp 4276-4285, August 2009.

R. Aggarwal, C. E. Koksal, and P. Schniter "Performance Bounds and Associated Design Principles for Multi-Cellular Wireless OFDMA Systems," *INFOCOM* 2012.

R. Aggarwal, M. Assaad, C. E. Koksal, and P. Schniter "Optimal Resource Allocation in OFDMA Downlink Systems With Imperfect CSI," *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications*, (San Francisco, CA), June 2011.

R. Aggarwal, M. Assaad, C. E. Koksal, and P. Schniter "OFDMA Downlink Resource Allocation via ARQ Feedback," *Proc. Asilomar Conf. on Signals, Systems, and Computers* (Pacific Grove, CA), Nov. 2009.

R. Aggarwal, P. Schniter, and C. E. Koksal "Rate Adaptation via ARQ Feedback for Goodput Maximization over Time-Varying Channels," *IEEE Conference on Information Sciences and Systems*, Princeton University, 2008.

Fields of Study

Major Field: Electrical and Computer Engineering

Table of Contents

Page

List of Tables

List of Figures

Chapter 1: Introduction

A wireless communication system is a system that enables communication between two or more users (or people or devices). Examples of such systems include mobile communication systems, satellite communication systems (GPS), AM/FM radio systems, and under-water communication systems. The most common wireless communication system in present-day world is the mobile communication system which supports communication between users having wireless devices, such as smartphones, tablets, and computers, via a network of service-nodes, such as base-stations, femtocells, and relays. In this dissertation, we will explore some issues that arise in the design of such systems.

Over the last decade, there has been a tremendous growth in the usage of mobile communication systems for information (or, data) transfer, typically in the form of voice-data or web-based data. This growth is attributed to the rise in number of wireless devices, particularly smartphones, and an increase in the number of web-based services, for example, video streaming, cloud-computing, banking etc. To meet this increasing data-demand of a growing number of users, better communication systems need to be designed. A complete study of a such big systems is quite complicated and is out of the scope of this dissertation. Here, we will focus on one of its aspects, namely

"resource allocation", using which efficiency/performance of the system can be enhanced. Examples of some other methods that are useful in supporting the increasing data-demand of users include developing better application-specific protocols, better data-compression algorithms, and better channel-coding schemes.

From a systems standpoint, the fundamental basis of any wireless communication system is the wireless link between transmitter(s) and receiver(s) that supports data transfer from transmitter(s) to receiver(s). The amount of information that can be transferred from transmitter(s) to receiver(s) over this link is limited by the amount of resources available at the transmitter(s). Typically, these resources are power and bandwidth. Further, these resources (power and bandwidth) are limited in almost every communication system. Therefore, efficient resource allocation schemes must be developed to exploit available resources in the best possible manner and provide ubiquitous high-data-rate to all users in a fair manner. In this dissertation, we study three different types of mobile-communication systems (point-to-point, single-cell OFDMA, and multi-cell OFDMA), and propose various optimal and/or near-optimal resource allocation schemes for each system. Using our analyses, we also provide design guidelines for service providers to design large multi-cellular communication systems while satisfying Quality-of-Service (QoS) requirements of users and revenue-targets of service providers.

1.1 Attacked Problems and Our Contributions

Brief descriptions of the problems considered in this dissertation, along with a summary of contributions towards each problem, are listed below:

1. In Chapter 2, we study the problem of code-rate adaptation in point-to-point links. In particular, we consider packetized time-varying continuous-state channels with link-layer feedback available at the transmitter in the form of per-packet ACK/NAKs (acknowledgements/negative acknowledgements). We propose a greedy code-rate adaptation algorithm and analyze the impact of limited (link-layer) feedback on per-packet information transfer on the "adaptiveness" of code-rate controllers. The metric associated with the performance is goodput. The sequence of ACK/NAKs gives a distributional estimate of the channel state. We use this estimate to achieve rate adaptation to maximize goodput.

   From the viewpoint of resource allocation, ACK/NAK feedback is quite different than pilot-aided feedback. While pilot-aided feedback provides an indicator of the *absolute* channel gain, ACK/NAK provides an indicator of the channel gain *relative* to the chosen user/power/rate; an ACK implies that the channel was good enough to support the user (with allocated power and rate) while a NAK implies otherwise. As a result, with ACK/NAK feedback based resource allocation, the chosen user/power/rate affects not only the subsequent utility but also the quality of the subsequent feedback, which in turn will affect future utilities through future resource assignments. In fact, with ACK/NAK feedback, optimal resource allocation assignment for communication over Markov channels turns out to be a POMDP policy with uncountable number of states and actions, which is impractical to implement.

   We show, via simulations, that the proposed rate adaptation scheme is a considerable improvement over fixed code-rate schemes with minimal extra computation. In fact, our scheme achieves up to 90% of the goodput gain achieved by

optimal genie-based schemes that use perfect causal/non-causal CSI for rate-adaptation over fixed code-rate schemes. We also show that the performance of the proposed (greedy) scheme outperforms the POMDP (Partially Observable Markov Decision Process) based optimal rate adaptation under discretized channel-models with up to 7 discrete states. Since POMDP-based optimal rate-adaptation for discrete-Markov channels with 7 or more states is computationally intensive, our greedy rate-adaptation scheme based on a continuous-Markov channel model is more appealing.

2. In Chapter 3, we study the downlink of a single-cell OFDMA (Orthogonal Frequency Division - Multiple Access) system under the availability of probabilistic CSI for all users at the transmitter (or, base-station). In cases where subchannel-sharing among users is allowed, we propose an optimal algorithm to simultaneously schedule users across OFDM subchannels, and allocate them powers and code-rates for system-wide utility maximization. In other cases where subchannel-sharing is not allowed, we propose a near-optimal algorithm and bound its performance. Our algorithms are bisection-based and are faster than other state-of-the-art algorithms (subgradient or golden-section based algorithms) that address resource allocation problems in OFDMA systems. Moreover, unlike other algorithms, theoretical performance guarantees as a function of any finite number of iterations are provided for our algorithms.

3. In Chapter 4, we consider an application of the previously considered joint scheduling and resource-allocation problem for utility maximization in single-cell OFDMA downlink systems. Here, binary ACK/NAKs are used to obtain

channel-state information (CSI) of the scheduled users at each time-slot. We propose a sub-optimal greedy resource allocation algorithm to schedule users, and allocate powers and code-rates. We show that a significant portion of the goodput gain achieved by the perfect-CSI based optimal algorithm over no-feedback can be achieved by our algorithm, which assumes: 1) usage of ACK/NAK feedback that is coarse, and 2) a situation where users that are not scheduled in a given time-slot do not send any feedback to the transmitting base-station. Our algorithm infers distributional estimates of channel-gains of every user using the available feedback (ACK/NAK) and uses them for scheduling. In order to compute these distributional estimates at each time-slot, we provide a computationally-efficient iterative algorithm based on particle filters.

4. In Chapter 5, we study multi-cellular OFDMA-based downlink systems and propose, for a general spatial geometry of transmitters and end-users, bounds on the achievable sum-rate as a function of the number, K, of users, the number, B, of base-stations, and the number, N, of available resource-blocks. Here, a resource block is a collection of subcarriers such that all such disjoint collections have associated independently fading channels. We evaluate the bounds for dense networks and regular-extended networks under uniform spatial distribution of users using *extreme-value theory*, and derive scaling laws for a truncated path-loss model and a variety of fading models (Rayleigh, Nakagami-m, Weibull, and LogNormal). We then provide design principles for the service providers that guarantee users' QoS constraints while maximizing revenue. Finally, we propose a practical scheme that achieves the same sum-rate scaling

law as that achieved by the optimal resource allocation policy for a wide range of parameters (K, B, N).

5. Chapter 6 gives conclusions, contributions, and directions for future work.

Proofs of various results in Chapters 2, 3, and 5 are provided in Appendix A, Appendix B, and Appendix C, respectively.

Chapter 2: Rate Adaptation via Link-Layer Feedback in Point-to-Point Links

2.1 Motivation

In point-to-point links, one transmitter wirelessly transmits data to one receiver. As mentioned in the previous chapter, a limited amount of resources, i.e, power and bandwidth, are available at the transmitter. Under an instantaneous power constraint, to maximize the reliability and amount of transmitted data, the transmitter will use all available power and bandwidth in each channel use. However, even in such cases, the maximum achievable goodput, defined as the maximum amount of data that can be transmitted on average, after discounting errors, may not be achieved. This is because most modulation and coding schemes that are used in practical systems are not capacity-achieving. Therefore, we are motivated to developed rate-adaptation schemes for point-to-point links in the presence of practical modulation and coding schemes that enable higher-goodput achievability.

Rate adaptation [1–15] achieves higher goodput by combating channel variability, which is common to all wireless communication systems due to factors such as fading, mobility, and multiuser interference. The idea is that, based on the predicted channel state, the transmitter optimizes the data rate in an effort to maximize the goodput.

For example, when the channel quality is below average, the data rate should be decreased to avoid reception errors, while, when the channel quality is above average, the data rate should be increased to prevent the channel from being underutilized.

Rate adaptation would be relatively straightforward if the transmitter could perfectly predict the channel. In practice, however, maintaining accurate transmitter channel state information is a nontrivial task that can consume valuable resources. Fortunately, error rate feedback in the form of single-bit ACK/NAKs are a standard provision in most networks by means of ARQ (Automatic Repeat reQuest). Therefore, rate adaptation using single-bit error-rate feedback comes at zero additional load on downlink/uplink channels and are of interest.

2.2 Overview and Related Work

In this chapter, we focus on rate adaptation schemes to improve performance of point-to-point systems wherein channel state knowledge is inferred by monitoring packet acknowledgments/negative-acknowledgments (ACK/NAKs) [5–15], i.e., the feedback information used for automatic repeat request (ARQ). Since ARQ feedback is a standard provision of the link layer, its use by the physical layer comes essentially "for free."

From the viewpoint of rate adaptation, ACK/NAK feedback is quite different than channel-state feedback. While channel-state feedback provides an indicator of the *absolute* channel gain, ACK/NAK provides an indicator of the channel gain *relative* to the chosen data rate; an ACK implies that the channel was good enough to support the rate while a NAK implies otherwise. As a result, with ACK/NAK-feedback based rate adaptation, the chosen data rate affects not only the subsequent goodput

but also the quality of the subsequent feedback, which in turn will affect future goodputs through future rate assignments. In fact, with ACK/NAK feedback, optimal rate assignment for communication over Markov channels can be recognized as a dynamic program [14], in particular, a partially observable Markov decision process (POMDP) [16].

In the next few sections, we consider the general problem of adapting the transmission rate via delayed and degraded error-rate feedback (in particular, ACK/NAK feedback) in order to maximize long-term expected goodput. In order to circumvent the sub-optimality of finite-state channel approximations [17], we assume a Markov channel indexed by a *continuous* parameter. Because the optimal solution of the POMDP is too difficult to obtain, we consider the use of *greedy* rate adaptation. First we establish that the optimal rate assignment is itself greedy when the error-rate feedback is *not* degraded. Furthermore, we establish that the greedy non-degraded scheme can be used to upper bound the optimal degraded scheme in terms of long-term goodput. Second, we outline a novel implementation of the greedy rate assignment scheme. For the example case of binary (i.e., ACK/NAK) degraded error-rate feedback, a Rayleigh-fading channel, and uncoded QAM modulation, we show (numerically) that the long-term goodput achieved by our greedy rate assignment scheme is close to the upper bound.

Compared to the previous works [5–13], which are ad hoc in nature, we take a more principled approach to cross-layer rate adaptation. Compared to the POMDP-based work [14], our work differs in the following key aspects: 1) we employ a continuous-state Markov channel model, 2) we consider delay in the feedback channel, and 3) we propose simpler greedy heuristics, which we study analytically as well as numerically.

Though our adaptation objective—goodput maximization—does not explicitly consider the input buffer state,[1] as does the one in [14], we show (numerically) that finite buffer effects (e.g., packet delay and drop rate) are handled gracefully by our greedy algorithms. In fact, one could argue that, since only successfully communicated packets are removed from the input queue, the maximization of *short-term* goodput—our greedy objective—leads simultaneously to the minimization of buffer occupancy.

This chapter is organized as follows. In Section 2.3, we outline our system model, and in Section 2.4, we consider optimal rate adaptation and suboptimal greedy approaches. In Section 2.5, we detail a novel implementation of the greedy rate-assignment scheme, which we then analyze numerically in Section 2.6 for the case of uncoded QAM transmission, ACK/NAK feedback, and a Rayleigh-fading channel. We summarize our findings in Section 2.7.

2.3 System Model

We consider a packetized transmission system in which the transmitter receives delayed and degraded feedback on the success of previous packet transmissions (e.g., binary ACK/NAKs), which it uses to adapt the subsequent transmission parameters. In particular, we assume the use of a transmission scheme parameterized by a data rate of r_t bits per packet, where t denotes the packet index. For simplicity, we assume a fixed transmission power and a fixed packet length of p channel uses.

Figure 2.1 shows the system model. The time-varying wireless channel is modeled by an SNR process $\{\gamma_t\}$, where the SNR $\gamma_t \geqslant 0$ is assumed to be constant over the packet duration. Notice that γ_t is not assumed to be a discrete parameter. Since the

[1] For algorithm design, we assume an infinitely back-logged queue.

transmission power is fixed, $\{\gamma_t\}$ is an exogenous process that does not depend on the transmission parameters. The instantaneous packet error rate $\epsilon(r_t, \gamma_t)$ varies with the rate r_t and SNR γ_t according to the particular modulation/demodulation scheme in use. The instantaneous goodput $G(r_t, \gamma_t)$, defined as the number of successfully communicated bits per channel use, is then defined as

$$G(r_t, \gamma_t) \triangleq \big(1 - \epsilon(r_t, \gamma_t)\big) r_t, \tag{2.1}$$

where $\epsilon(r_t, \gamma_t)$ denotes the error rate at time t. The transmitter uses $\hat{\epsilon}_{t-d}$, an estimated version of the ($d \geqslant 1$ delayed) error rate $\epsilon(r_{t-d}, \gamma_{t-d})$ to choose the time-t rate parameter r_t. We will assume that, for each r_t, the function $\epsilon(r_t, \gamma_t)$ is monotonically decreasing in γ_t, so that γ_t can be uniquely determined given r_t and the true error rate $\epsilon(r_t, \gamma_t)$.

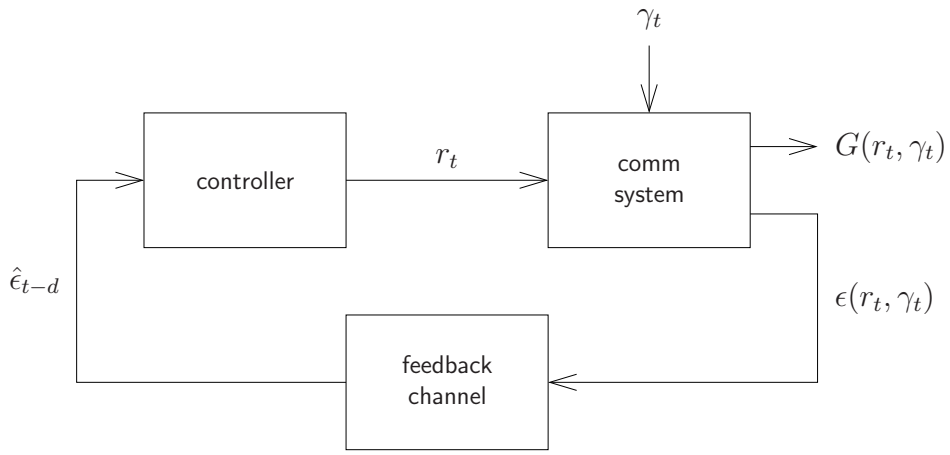

Figure 2.1: System model.

Example 1. *As an illustrative example, we now consider uncoded quadrature amplitude modulation (QAM) using a square constellation of size m, and minimum-distance*

decision making. At the link layer, where symbols are grouped to form packets, a fixed number of extra cyclic redundancy check (CRC) bits are appended to each packet for the purpose of error detection. We will assume the probability of undetected error is negligible and the associated ACK/NAK error feedback is sent back to the transmitter over an error-free reverse channel. We will also assume that the number of CRC bits is small compared to the packet size, allowing us to ignore them in goodput calculations.

Under an AWGN channel, and with γ describing the ratio of received symbol power to additive noise power, the symbol error rate for minimum-distance decision making is [18, p. 280]

$$1 - \left(1 - 2\left(1 - \frac{1}{\sqrt{m}}\right) Q\left(\sqrt{\frac{3\gamma}{m-1}}\right)\right)^2, \tag{2.2}$$

where $Q(\cdot)$ denotes the Q-function [18]. If we assume that the constellation size is fixed over the packet duration, then the data rate equals $r_t = p \log_2 m_t$ and the packet error rate equals

$$\epsilon(r_t, \gamma_t) \;=\; 1 - \left(1 - 2\left(1 - \frac{1}{\sqrt{2^{r_t/p}}}\right) Q\left(\sqrt{\frac{3\gamma_t}{2^{r_t/p} - 1}}\right)\right)^{2p}. \tag{2.3}$$

Plugging (2.3) into (2.1) yields the instantaneous goodput expression, which identifies a particular one-to-one mapping between rate r_t and goodput for a fixed SNR γ_t. Thus, if the SNR was known perfectly, then the goodput could be maximized by appropriate choice of constellation size. Figure 2.2 plots instantaneous goodput contours versus SNR γ_t and constellation size m_t for the case of $p = 100$ symbols per packet. Figure 2.2 also plots the (unique) goodput-maximizing constellation size as a function of SNR. Here the finite set of allowed constellation choices (and hence rates) is apparent. Note that, for this uncoded communication scheme, the SNR must be relatively high

to facilitate rate adaptation; as long as the SNR remains below 14 *dB, the goodput-maximizing constellation size remains at* $m = 4$ *(i.e., QPSK). Coded transmission, on the other hand, could facilitate rate adaptation at lower SNRs.*

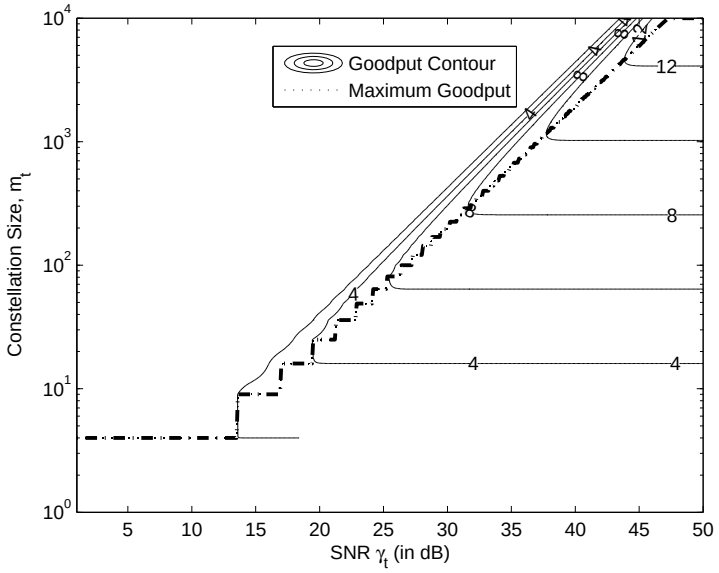

Figure 2.2: Goodput contours versus SNR γ_t and constellation size m_t for packet size $p = 100$. The goodput maximizing constellation size, as a function of SNR, is shown by the dash-dot line.

With ACK/NAK error feedback, the estimated error-rate $\hat{\epsilon}_t$ *is a Bernoulli random variable generated from* $\epsilon(r_t, \gamma_t)$ *according to the conditional probability mass function*

$$p\big(\hat{\epsilon}_t = k \big| \epsilon(r_t, \gamma_t)\big) \quad = \quad \begin{cases} \epsilon(r_t, \gamma_t) & k = 1 \\ 1 - \epsilon(r_t, \gamma_t) & k = 0 \\ 0 & \textit{else.} \end{cases} \tag{2.4}$$

□

While the previous example focuses on a particular modulation/demodulation scheme and a particular error feedback model, we emphasize that the principal results

in the sequel are general; no particular modulation/demodulation scheme and error feedback model are assumed.

2.4 Optimal Rate Adaptation

In this section we formalize the problem of finite horizon goodput maximization. For convenience we assume that process $\{\gamma_t, \hat{\epsilon}_t, r_t\}$ has been initiated at time $t = -\infty$, though we consider only the finite sequence of packet indices $\{0, \ldots, T\}$ for goodput maximization. Also, we use the abbreviation $\epsilon_t \triangleq \epsilon(r_t, \gamma_t)$.

For every packet index $t \geqslant 0$, we assume that the rate controller has access to the estimated error-rate feedback $\hat{\boldsymbol{\epsilon}}_{t-d} \triangleq [\ldots, \hat{\epsilon}_{-2}, \hat{\epsilon}_{-1}, \hat{\epsilon}_0, \ldots, \hat{\epsilon}_{t-d}]$, where $d \geqslant 1$ denotes the causal feedback delay. Formally, we consider $\hat{\boldsymbol{\epsilon}}_{t-d}$ to be *degraded* relative to the true error-rate vector ϵ_{t-d} if

$$\begin{aligned} \mathrm{E}\{G(r_t, \gamma_t) \mid \hat{\boldsymbol{\epsilon}}_{t-d}, \mathbf{r}_{t-d}\} &\neq \mathrm{E}\{G(r_t, \gamma_t) \mid \epsilon_{t-d}, \mathbf{r}_{t-d}\} && (2.5) \\ &= \mathrm{E}\{G(r_t, \gamma_t) \mid \gamma_{t-d}\}, && (2.6) \end{aligned}$$

where $\mathbf{r}_{t-d} \triangleq [\ldots, r_{-2}, r_{-1}, r_0, \ldots, r_{t-d}]$. Equation (2.6) follows because each SNR γ_k in $\gamma_{t-d} \triangleq [\ldots, \gamma_{-2}, \gamma_{-1}, \gamma_0, \ldots, \gamma_{t-d}]$ can be uniquely determined from the pair (ϵ_k, r_k).

At packet index t, the *optimal* controller uses the degraded error-rate sequence $\hat{\boldsymbol{\epsilon}}_{t-d}$ (as well as knowledge of the previously chosen rates $\mathbf{r}_{t-d}$) to choose the rate r_t from a set $\mathcal{R}$ of admissible rates in order to maximize the total expected goodput for the current and remaining packets:

$$r_t^* \triangleq \arg\max_{r_t \in \mathcal{R}} \mathrm{E}\left\{ G(r_t, \gamma_t) + \sum_{k=t+1}^{T} G(r_k^*, \gamma_k) \;\middle|\; \hat{\boldsymbol{\epsilon}}_{t-d}, \mathbf{r}_{t-d} \right\} \quad \text{for} \quad t = 0, \ldots, T \qquad (2.7)$$

The optimal expected sum goodput for packets $\{t, \ldots, T\}$ can then be written (for $t \geqslant 0$) as

$$G_t^*(\hat{\boldsymbol{\epsilon}}_{t-d}, \mathbf{r}_{t-d}) \triangleq \mathrm{E}\left\{ \sum_{k=t}^{T} G(r_k^*, \gamma_k) \;\middle|\; \hat{\boldsymbol{\epsilon}}_{t-d}, \mathbf{r}_{t-d} \right\}. \tag{2.8}$$

For a unit[2] delay system (i.e., $d = 1$), the following Bellman equation [19] specifies the associated finite-horizon dynamic programming problem:

$$\begin{aligned} G_t^*(\hat{\boldsymbol{\epsilon}}_{t-1}, \mathbf{r}_{t-1}) &= \max_{r_t \in \mathcal{R}} \big\{ \mathrm{E}\{G(r_t, \gamma_t) \mid \hat{\boldsymbol{\epsilon}}_{t-1}, \mathbf{r}_{t-1}\} \\ &\qquad + \mathrm{E}\{G_{t+1}^*([\hat{\boldsymbol{\epsilon}}_{t-1}, \hat{\epsilon}_t], [\mathbf{r}_{t-1}, r_t]) \mid \hat{\boldsymbol{\epsilon}}_{t-1}, \mathbf{r}_{t-1}\}\big\}, \end{aligned} \tag{2.9}$$

where the second expectation is over $\hat{\epsilon}_t$. The solution to this problem is sometimes referred to as a partially observable Markov decision process (POMDP) [16].

For practical horizons T, optimal rate selection based on (2.9) is intractable, in part due to the continuous-state nature of the channel.[3] In fact, it is known that POMDPs are PSPACE-complete, i.e., they require both complexity and memory that grow exponentially with the horizon T [20]. For an intuitive understanding of this phenomenon, notice from (2.9) that the solution of the rate assignment problem at every time t depends on the optimal rate assignments up to time $t-1$. But, because both terms on the right side of (2.9) are dependent on r_t, the solution of the rate assignment problem at time t also depends on the solution of the rate assignment problem at time $t+1$, which in turn depends on the solution of the rate assignment problem at time $t+2$, and so on. Consequently, the much simpler *greedy*

[2] For the $d > 1$ case, the Bellman equation is more complicated, and so we omit it for brevity.

[3] Though a quantized channel approximation could—with few enough states—yield a tractable POMDP solution, we show in Section 2.6 that channel quantization leads to significant loss in goodput.

rate assignment scheme

$$\hat{r}_t \triangleq \arg\max_{r_t \in \mathcal{R}} \mathrm{E}\{G(r_t, \gamma_t) \mid \hat{\boldsymbol{\epsilon}}_{t-d}, \mathbf{r}_{t-d}\} \text{ for } t = 0, \ldots, T, \tag{2.10}$$

is suboptimal.

The question of principal interest is then: *What is the loss in goodput with the greedy scheme (2.10) relative to the optimal scheme (2.7)?* Since it is too difficult to compute the optimal goodput (which depends on the optimal rate assignment $\mathbf{r}_T^*$), we instead compare the greedy scheme (2.10) to an *upper bound* on the optimal goodput. To establish the upper bound, we show that greedy rate assignment using *non-degraded* error-rate feedback yields a total goodput that is no less than that of optimal rate assignment using degraded error-rate feedback. While the latter is difficult to compute, the former is not.

We now detail the rate assignment scheme that leads to our total-goodput upper bound. At packet index t, consider the rate assignment that maximizes the total expected goodput for the current and remaining packets using knowledge of the non-degraded feedback ϵ_{t-d}:

$$r_t^{\mathrm{cg}} \triangleq \arg\max_{r_t \in \mathcal{R}} \mathrm{E}\left\{ G(r_t, \gamma_t) + \sum_{k=t+1}^{T} G(r_k^{\mathrm{cg}}, \gamma_k) \;\middle|\; \epsilon_{t-d}, \mathbf{r}_{t-d} \right\} \text{ for } t = 0, \ldots, T. \tag{2.11}$$

We refer to this scheme as the *causal genie*. Note that (2.11) differs from (2.7) only in that ϵ_{t-d} is used in place of $\hat{\boldsymbol{\epsilon}}_{t-d}$. Because γ_{t-d} can be uniquely determined from $(\epsilon_{t-d}, \mathbf{r}_{t-d})$, the causal genie can also be written as

$$r_t^{\mathrm{cg}} = \arg\max_{r_t \in \mathcal{R}} \mathrm{E}\left\{ G(r_t, \gamma_t) + \sum_{k=t+1}^{T} G(r_k^{\mathrm{cg}}, \gamma_k) \;\middle|\; \gamma_{t-d} \right\} \text{ for } t = 0, \ldots, T. \tag{2.12}$$

Since the choice of $\{r_k^{\text{cg}}\}_{k=t+1}^{T}$ will not depend on the choice of r_t, the optimal expected sum goodput for packets $\{t, \ldots, T\}$ can be written (for $t \geqslant 0$) as

$$G_t^{\text{cg}}(\epsilon_{t-d}, \mathbf{r}_{t-d}) \triangleq \max_{r_t \in \mathcal{R}} \mathrm{E}\left\{ G(r_t, \gamma_t) + \sum_{k=t+1}^{T} G(r_k^{\text{cg}}, \gamma_k) \;\middle|\; \gamma_{t-d} \right\} \tag{2.13}$$

$$= \mathrm{E}\left\{ \sum_{k=t+1}^{T} G(r_k^{\text{cg}}, \gamma_k) \;\middle|\; \gamma_{t-d} \right\} + \max_{r_t \in \mathcal{R}} \mathrm{E}\{G(r_t, \gamma_t) \mid \gamma_{t-d}\}, \tag{2.14}$$

which shows that *optimal rate assignment under non-degraded causal error-rate feedback can be accomplished greedily.* In other words,

$$r_t^{\text{cg}} = \arg\max_{r_t \in \mathcal{R}} \mathrm{E}\{G(r_t, \gamma_t) \mid \gamma_{t-d}\} \tag{2.15}$$

$$= \arg\max_{r_t \in \mathcal{R}} \mathrm{E}\{G(r_t, \gamma_t) \mid \epsilon_{t-d}, \mathbf{r}_{t-d}\}. \tag{2.16}$$

We now establish that the causal genie controller upper bounds the optimal controller with degraded error-rate feedback in the sense of total goodput. Though the result may be intuitive, the proof provides insight into the relationship between degradation of the feedback and reduction of the total expected goodput.

Lemma 1. *Given arbitrary past rates $\mathbf{r}_{-d}$ and corresponding degraded error-rate feedback $\hat{\boldsymbol{\epsilon}}_{-d}$, the expected total goodput for optimal rate allocation under degraded feedback is no higher than the expected total goodput for the causal-genie rate allocation under non-degraded feedback, i.e.,*

$$\mathrm{E}\left\{ \sum_{t=0}^{T} G(r_t^*, \gamma_t) \;\middle|\; \hat{\boldsymbol{\epsilon}}_{-d}, \mathbf{r}_{-d} \right\} \leqslant \mathrm{E}\left\{ \sum_{t=0}^{T} G(r_t^{\text{cg}}, \gamma_t) \;\middle|\; \hat{\boldsymbol{\epsilon}}_{-d}, \mathbf{r}_{-d} \right\}. \tag{2.17}$$

Proof. For any $t \in \{0, \ldots, T\}$ and any realization of $(\hat{\boldsymbol{\epsilon}}_{t-d}, \mathbf{r}_{t-d})$, we can write

$$\begin{aligned}
\mathrm{E}\{G(r_t^*, \gamma_t) \mid \hat{\boldsymbol{\epsilon}}_{t-d}, \mathbf{r}_{t-d}\} &\leqslant \max_{r_t \in \mathcal{R}} \mathrm{E}\{G(r_t, \gamma_t) \mid \hat{\boldsymbol{\epsilon}}_{t-d}, \mathbf{r}_{t-d}\} && (2.18)\\
&= \max_{r_t \in \mathcal{R}} \mathrm{E}\left\{ \mathrm{E}\{G(r_t, \gamma_t) \mid \hat{\boldsymbol{\epsilon}}_{t-d}, \mathbf{r}_{t-d}, \epsilon_{t-d}\} \;\middle|\; \hat{\boldsymbol{\epsilon}}_{t-d}, \mathbf{r}_{t-d} \right\} && (2.19)\\
&\leqslant \mathrm{E}\left\{ \max_{r_t \in \mathcal{R}} \mathrm{E}\{G(r_t, \gamma_t) \mid \hat{\boldsymbol{\epsilon}}_{t-d}, \mathbf{r}_{t-d}, \epsilon_{t-d}\} \;\middle|\; \hat{\boldsymbol{\epsilon}}_{t-d}, \mathbf{r}_{t-d} \right\} && (2.20)\\
&= \mathrm{E}\left\{ \max_{r_t \in \mathcal{R}} \mathrm{E}\{G(r_t, \gamma_t) \mid \gamma_{t-d}\} \;\middle|\; \hat{\boldsymbol{\epsilon}}_{t-d}, \mathbf{r}_{t-d} \right\} && (2.21)\\
&= \mathrm{E}\{G(r_t^{\mathrm{cg}}, \gamma_t) \mid \hat{\boldsymbol{\epsilon}}_{t-d}, \mathbf{r}_{t-d}\}, && (2.22)
\end{aligned}$$

where (2.18) follows since r_t^* is chosen to maximize the long term goodput—not the instantaneous goodput; (2.20) follows since $\max_{r_t} \mathrm{E}\{f(r_t)\} \leqslant \mathrm{E}\{\max_{r_t} f(r_t)\}$ for any $f(\cdot)$; (2.21) follows by definition of degraded feedback; and (2.22) follows by definition of the greedy genie. Taking the expectation over $(\hat{\boldsymbol{\epsilon}}_{t-d}, \mathbf{r}_{t-d})$, conditional on $(\hat{\boldsymbol{\epsilon}}_{-d}, \mathbf{r}_{-d})$, we find

$$\mathrm{E}\{G(r_t^*, \gamma_t) \mid \hat{\boldsymbol{\epsilon}}_{-d}, \mathbf{r}_{-d}\} \leqslant \mathrm{E}\{G(r_t^{\mathrm{cg}}, \gamma_t) \mid \hat{\boldsymbol{\epsilon}}_{-d}, \mathbf{r}_{-d}\}. \quad (2.23)$$

Finally, summing both sides of (2.23) over $t = \{0, \ldots, T\}$ yields (2.17). □

In Section 2.5 we study the greedy rate assignment scheme (2.10) in depth. Then, in Section 2.6, we study (numerically) the particular case in which $\{\hat{\epsilon}_t\}_{t \geqslant 0}$ is constructed from link-layer ACK/NAKs.

2.5 The Greedy Rate Adaptation Algorithm

In this section, we detail the implementation of greedy rate assignment (2.10) assuming continuous Markov SNR variation and conditionally independent error-rate estimates. In Section 2.5.1, we detail a procedure for packet-rate adaptation, while in Section 2.5.2, we consider adapting the rate once per block of n packets.

2.5.1 Packet-Rate Algorithm

Assuming a feedback delay of $d \geqslant 1$ packets, the greedy rate assignment (2.10) can be rewritten as

$$\hat{r}_t \;=\; \arg\max_{r_t \in \mathcal{R}} \int G(r_t, \gamma_t) p(\gamma_t \mid \hat{\boldsymbol{\epsilon}}_{t-d}, \mathbf{r}_{t-d})\; d\gamma_t \quad \text{for } t = 0, \ldots, T. \tag{2.24}$$

We now derive a recursive implementation of the greedy rate assignment (2.24).

Expanding the inferred SNR distribution via Bayes rule, we find

$$\begin{aligned} & p(\gamma_t \mid \hat{\epsilon}_{t-d}, \mathbf{r}_{t-d}) \\ &= \int p(\gamma_t \mid \gamma_{t-d}, \hat{\epsilon}_{t-d}, \mathbf{r}_{t-d}) p(\gamma_{t-d} \mid \hat{\epsilon}_{t-d}, \mathbf{r}_{t-d})\; d\gamma_{t-d} && (2.25) \\ &= \int p(\gamma_t \mid \gamma_{t-d}) p(\gamma_{t-d} \mid \hat{\epsilon}_{t-d}, \mathbf{r}_{t-d})\; d\gamma_{t-d}, && (2.26) \end{aligned}$$

where we used the assumption of Markov SNR variation to write (2.26). Furthermore,

$$\begin{aligned} & p(\gamma_{t-d} \mid \hat{\epsilon}_{t-d}, \mathbf{r}_{t-d}) \\ &= p(\gamma_{t-d} \mid \hat{\epsilon}_{t-d}, \hat{\epsilon}_{t-d-1}, \mathbf{r}_{t-d}) && (2.27) \\ &= \frac{p(\hat{\epsilon}_{t-d} \mid \gamma_{t-d}, \hat{\epsilon}_{t-d-1}, \mathbf{r}_{t-d}) p(\gamma_{t-d} \mid \hat{\epsilon}_{t-d-1}, \mathbf{r}_{t-d})}{\int p(\hat{\epsilon}_{t-d} \mid \gamma'_{t-d}, \hat{\epsilon}_{t-d-1}, \mathbf{r}_{t-d}) p(\gamma'_{t-d} \mid \hat{\epsilon}_{t-d-1}, \mathbf{r}_{t-d})\; d\gamma'_{t-d}} && (2.28) \\ &= \frac{p(\hat{\epsilon}_{t-d} \mid \epsilon(r_{t-d}, \gamma_{t-d}), \hat{\epsilon}_{t-d-1}) p(\gamma_{t-d} \mid \hat{\epsilon}_{t-d-1}, \mathbf{r}_{t-d-1})}{\int p(\hat{\epsilon}_{t-d} \mid \epsilon(r_{t-d}, \gamma'_{t-d}), \hat{\epsilon}_{t-d-1}) p(\gamma'_{t-d} \mid \hat{\epsilon}_{t-d-1}, \mathbf{r}_{t-d-1})\; d\gamma'_{t-d}}. && (2.29) \end{aligned}$$

With conditionally independent error estimates (i.e., $p(\hat{\epsilon}_t \mid \epsilon_t, \hat{\boldsymbol{\epsilon}}_{t-1}) = p(\hat{\epsilon}_t \mid \epsilon_t)$), this becomes

$$\begin{aligned} & p(\gamma_{t-d} \mid \hat{\epsilon}_{t-d}, \mathbf{r}_{t-d}) \\ &= \frac{p(\hat{\epsilon}_{t-d} \mid \epsilon(r_{t-d}, \gamma_{t-d})) p(\gamma_{t-d} \mid \hat{\epsilon}_{t-d-1}, \mathbf{r}_{t-d-1})}{\int p(\hat{\epsilon}_{t-d} \mid \epsilon(r_{t-d}, \gamma'_{t-d})) p(\gamma'_{t-d} \mid \hat{\epsilon}_{t-d-1}, \mathbf{r}_{t-d-1})\; d\gamma'_{t-d}}. && (2.30) \end{aligned}$$

Similar to (2.26), we can also write

$$\begin{aligned} & p(\gamma_{t-d+1} \mid \hat{\epsilon}_{t-d}, \mathbf{r}_{t-d}) \\ &= \int p(\gamma_{t-d+1} \mid \gamma_{t-d}, \hat{\epsilon}_{t-d}, \mathbf{r}_{t-d}) \\ &\quad \times p(\gamma_{t-d} \mid \hat{\epsilon}_{t-d}, \mathbf{r}_{t-d})\, d\gamma_{t-d} \qquad (2.31) \\ &= \int p(\gamma_{t-d+1} \mid \gamma_{t-d}) p(\gamma_{t-d} \mid \hat{\epsilon}_{t-d}, \mathbf{r}_{t-d})\, d\gamma_{t-d}. \qquad (2.32) \end{aligned}$$

Equations (2.26), (2.30), and (2.32) lead to the following recursive implementation of the greedy rate assignment (2.24). Assuming the availability[4] of $p(\gamma_{t-d} \mid \hat{\epsilon}_{t-d-1}, \mathbf{r}_{t-d-1})$ when calculating r_t, the rate assignment procedure for packet indices $t = 0, \ldots, T$ is:

1. Measure $\hat{\epsilon}_{t-d}$, compute $p(\hat{\epsilon}_{t-d} \mid \epsilon(r_{t-d}, \gamma_{t-d}))$ as a function of γ_{t-d}, and then calculate the distribution $p(\gamma_{t-d} \mid \hat{\boldsymbol{\epsilon}}_{t-d}, \mathbf{r}_{t-d})$ using (2.30).

2. Calculate $p(\gamma_t \mid \hat{\boldsymbol{\epsilon}}_{t-d}, \mathbf{r}_{t-d})$ using the Markov prediction step (2.26).

3. Calculate $\hat{r}_t$ via (2.24).

4. If[5] $d > 1$, then calculate $p(\gamma_{t-d+1} \mid \hat{\epsilon}_{t-d}, \mathbf{r}_{t-d})$ via (2.32) for use in the next iteration.

2.5.2 Block-Rate Algorithm

Since it may be impractical for the transmitter to adapt the rate on a per-packet basis, we now propose a modification of the algorithm detailed in Section 2.5.1 that adapts the rate only once per block of n packets. The main idea behind our block-rate algorithm is that the SNR $\{\gamma_t\}$ and error-rate estimates $\{\hat{\epsilon}_t\}$ are treated as if they

[4] For the initial packet indices $t \in \{0, \ldots, d\}$, if the pdf $p(\gamma_{t-d} \mid \hat{\epsilon}_{t-d-1}, \mathbf{r}_{t-d-1})$ is unknown, then we suggest to use the prior $p(\gamma_{t-d})$ in its place.

[5] Notice that, if $d = 1$, then $p(\gamma_{t-d+1} \mid \hat{\epsilon}_{t-d}, \mathbf{r}_{t-d})$ was already computed in step 2).

were *constant* over the block, thereby allowing a straightforward application of the method from Section 2.5.1. Though this treatment is suboptimal, our intention is to trade performance for reduced complexity.

The details of our block-rate algorithm are now given. Denoting the block index by i, the block versions of the degraded error-rate estimate and SNR are defined as

$$\underline{\hat{\epsilon}}_i \triangleq \frac{1}{n} \sum_{t=in}^{(i+1)n-1} \hat{\epsilon}_t \tag{2.33}$$

$$\underline{\gamma}_i \triangleq \gamma_{in+\lfloor n/2 \rfloor}, \tag{2.34}$$

and the assigned rates $\{r_t\}$ are related to the calculated rates $\{\underline{r}_i\}$ as

$$r_t = \underline{r}_{\lfloor t/n \rfloor}. \tag{2.35}$$

Notice that, when $n = 1$, the block-rate quantities reduce to the packet-rate quantities, i.e., $\underline{\hat{\epsilon}}_i = \hat{\epsilon}_i$, $\underline{\gamma}_i = \gamma_i$, and $\underline{r}_i = r_i$.

Borrowing the packet-rate adaptation approach from Section 2.5.1, the block-rate greedy implementation goes as follows. Here, we use d to denote the delay in *blocks*. Assuming the availability[6] of $p(\underline{\gamma}_{i-d} \mid \underline{\hat{\boldsymbol{\epsilon}}}_{i-d-1}, \underline{\boldsymbol{r}}_{i-d-1})$ when calculating $\underline{r}_i$, the rate assignment procedure for block indices $i = 0, \ldots, \lceil T/n \rceil$ is:

1. Measure $\{\hat{\epsilon}_t\}_{t=(i-d)n}^{(i-d+1)n-1}$, compute $\underline{\hat{\epsilon}}_{i-d}$ via (2.33), compute $p(\underline{\hat{\epsilon}}_{i-d} \mid \epsilon(\underline{r}_{i-d}, \underline{\gamma}_{i-d}))$ as a function of $\underline{\gamma}_{i-d}$, and then calculate the inferred SNR distribution $p(\underline{\gamma}_{i-d} \mid \underline{\hat{\boldsymbol{\epsilon}}}_{i-d}, \underline{\boldsymbol{r}}_{i-d})$ using

$$p(\underline{\gamma}_{i-d} \mid \underline{\hat{\boldsymbol{\epsilon}}}_{i-d}, \underline{\boldsymbol{r}}_{i-d}) = \frac{p(\underline{\hat{\epsilon}}_{i-d} \mid \epsilon(\underline{r}_{i-d}, \underline{\gamma}_{i-d}))p(\underline{\gamma}_{i-d} \mid \underline{\hat{\boldsymbol{\epsilon}}}_{i-d-1}, \underline{\boldsymbol{r}}_{i-d-1})}{\int p(\underline{\hat{\epsilon}}_{i-d} \mid \epsilon(\underline{r}_{i-d}, \underline{\gamma}'_{i-d}))p(\underline{\gamma}'_{i-d} \mid \underline{\hat{\boldsymbol{\epsilon}}}_{i-d-1}, \underline{\boldsymbol{r}}_{i-d-1})\, d\underline{\gamma}'_{i-d}}. \tag{2.36}$$

[6] For the initial block indices $i \in \{0, \ldots, d\}$, if the pdf $p(\underline{\gamma}_{i-d} \mid \underline{\hat{\boldsymbol{\epsilon}}}_{i-d-1}, \underline{\boldsymbol{r}}_{i-d-1})$ is unknown, then we suggest to use the prior $p(\underline{\gamma}_{i-d})$ in its place.

2. Calculate $p(\underline{\gamma}_i \mid \hat{\underline{\boldsymbol{\epsilon}}}_{i-d}, \underline{\boldsymbol{r}}_{i-d})$ using the Markov prediction step

$$\begin{aligned} & p(\underline{\gamma}_i \mid \hat{\underline{\boldsymbol{\epsilon}}}_{i-d}, \underline{\boldsymbol{r}}_{i-d}) \\ &= \int p(\underline{\gamma}_i \mid \underline{\gamma}_{i-d}) p(\underline{\gamma}_{i-d} \mid \hat{\underline{\boldsymbol{\epsilon}}}_{i-d}, \underline{\boldsymbol{r}}_{i-d}) \; d\underline{\gamma}_{i-d}. \end{aligned} \tag{2.37}$$

3. Calculate $\hat{\underline{r}}_i$ via

$$\hat{\underline{r}}_i = \arg\max_{\underline{r}_i \in \mathcal{R}} \int G(\underline{r}_i, \underline{\gamma}_i) p(\underline{\gamma}_i \mid \hat{\underline{\boldsymbol{\epsilon}}}_{i-d}, \underline{\boldsymbol{r}}_{i-d}) \; d\underline{\gamma}_i. \tag{2.38}$$

4. If[7] $d > 1$, then calculate $p(\underline{\gamma}_{i-d+1} \mid \hat{\underline{\boldsymbol{\epsilon}}}_{i-d}, \underline{\boldsymbol{r}}_{i-d})$ as follows for use in the next iteration.

$$\begin{aligned} & p(\underline{\gamma}_{i-d+1} \mid \hat{\underline{\boldsymbol{\epsilon}}}_{i-d}, \underline{\boldsymbol{r}}_{i-d}) \\ &= \int p(\underline{\gamma}_{i-d+1} \mid \underline{\gamma}_{i-d}) p(\underline{\gamma}_{i-d} \mid \hat{\underline{\boldsymbol{\epsilon}}}_{i-d}, \underline{\boldsymbol{r}}_{i-d}) \; d\underline{\gamma}_{i-d}. \end{aligned} \tag{2.39}$$

As the adaptation-block size n increases, we expect the packet error rate estimate $\hat{\underline{\epsilon}}_t$ to become more accurate (since it is estimated from, e.g., n ACK/NAKs), the SNR model to get less accurate (since a block-fading approximation is being applied to a process that is continuously fading), and the per-packet implementation complexity of the algorithm to decrease.

We note that the block-rate modification proposed here is suboptimal in the sense that the SNR of each packet in a block could have been predicted individually, rather than predicting only the SNR of the packet in the middle of the block. Likewise, individual rates could have been assigned for each packet in the block, rather than a uniform rate for all packets in the block. However, joint optimization of intra-block rates appears to be prohibitively complex and thus goes against our primary motivation for the block-rate algorithm, i.e., simplicity.

[7] Notice that, if $d = 1$, then $p(\underline{\gamma}_{i-d+1} \mid \hat{\underline{\boldsymbol{\epsilon}}}_{i-d}, \underline{\boldsymbol{r}}_{i-d})$ was already computed in step 2).

Finally, we note that a similar block-rate modification can also be applied to the causal genie scheme (2.16), which has been recognized as a non-degraded-feedback version of the greedy scheme (2.10). However, doing so would spoil the total-goodput optimality of the packet-rate causal genie that was identified in Lemma 1.

2.6 Numerical Results

We now describe the results of numerical experiments in which we assume uncoded square-QAM modulation, a Gauss-Markov fading channel, and minimum variance unbiased (MVU) estimation of the error-rate, as detailed below. While other examples of modulation, error-rate estimation, and fading could have been employed, we feel that our choices are sufficient to illustrate the essential behaviors of the generic rate adaptation schemes discussed in Sections 2.4, 2.5.

2.6.1 Setup

For our numerical experiments, we used the uncoded QAM modulation/demodulation scheme described in Example 1, which yields the packet error-rate given in (2.3). We used squared-integer constellation sizes, i.e., 4-QAM, 9-QAM, 16-QAM, etc. In addition, we used causal degraded error-rate feedback in the form of one ACK/NAK per transmitted packet. Thus, in a block[8] of n packets, there were n ACK/NAKs.

Given this setup, it can be shown that the MVU estimate [21] of the average packet error rate over the i-th block can be computed by a simple arithmetic average of the n ACK/NAKs, using 0 for an ACK and 1 for a NAK. Notice that this MVU estimate corresponds exactly to the block error-rate estimate $\hat{\underline{\epsilon}}_i$ specified in (2.33). Furthermore, if $\underline{\epsilon}_i$ denotes the value of the true packet error rate over the i-th block,

[8] The results here also hold for packet-rate adaptation through the choice $n = 1$.

then the number of NAKs per block is Binomial$(n, \underline{\epsilon}_i)$ and the error estimate $\hat{\underline{\epsilon}}_i$ obeys

$$p(\hat{\underline{\epsilon}}_i = \tfrac{k}{n} \mid \underline{\epsilon}_i) \;=\; \begin{cases} \binom{n}{k} \underline{\epsilon}_i^k (1-\underline{\epsilon}_i)^{n-k} & \text{for } k = 0, \ldots, n \\ 0 & \text{else.} \end{cases} \tag{2.40}$$

Thus, we can calculate $\underline{\epsilon}_i = \epsilon(\underline{r}_i, \underline{\gamma}_i)$ as a function of $\underline{\gamma}_i$ using (2.3) and plug the results into $p(\hat{\underline{\epsilon}}_i \mid \underline{\epsilon}_i)$ from (2.40) in order to compute (2.36).

To generate the Markov block-rate SNR process $\{\underline{\gamma}_i\}$, we first generate a packet-rate complex-valued Gauss-Markov "channel gain" [22] process $\{g_t\}$ using

$$g_t \;=\; (1-\alpha) g_{t-1} + \alpha w_t, \tag{2.41}$$

where $\{w_t\}$ is a zero-mean unit-variance white circular Gaussian driving process and $0 \leq \alpha \leq 1$. Notice that $\alpha = 1$ corresponds to i.i.d. gains, whereas $\alpha = 0$ corresponds to a time-invariant gain. We then generate a packet-rate SNR process $\{\gamma_t\}$ by scaling the squared magnitude of g_t:

$$\gamma_t \;=\; K|g_t|^2. \tag{2.42}$$

The scaling parameter K in (2.42) is essential because α affects both the (steady-state) coherence time and the mean-squared value of the gain $\{g_t\}$. Thus, by using the two parameters K and α, it is possible to independently control the (steady-state) mean and coherence time of the SNR process $\{\gamma_t\}$. In fact, it can be shown that, for steady-state indices t, the SNR γ_t is exponentially distributed with mean value $\frac{2K\alpha}{2-\alpha}$.

To evaluate $p(\underline{\gamma}_i \mid \underline{\gamma}_{i-d})$, we first notice from (2.34) that $p(\underline{\gamma}_i | \underline{\gamma}_{i-d}) = p(\gamma_t \mid \gamma_{t-nd})$. Then, from (2.41), we find that

$$g_t \;=\; (1-\alpha)^{nd} g_{t-nd} + \alpha \sum_{j=0}^{nd-1} (1-\alpha)^j w_{t-j}, \tag{2.43}$$

where $\sum_{j=0}^{nd-1}(1-\alpha)^j w_{t-j} \sim \mathcal{CN}\big(0, \frac{1}{1-(1-\alpha)^2}(1-(1-\alpha)^{2nd})\big)$. From this fact, we show in Appendix A that

$$\begin{aligned} p(\gamma_t \mid \gamma_{t-nd}) &= \frac{2-\alpha}{2K\alpha(1-(1-\alpha)^{2nd})} \\ &\quad \times \exp\left(\frac{-(\gamma_t + (1-\alpha)^{2nd}\gamma_{t-nd})(2-\alpha)}{2K\alpha(1-(1-\alpha)^{2nd})}\right) \\ &\quad \times I_o\left(\frac{(1-\alpha)^{nd}\sqrt{\gamma_t\gamma_{t-nd}}(2-\alpha)}{K\alpha(1-(1-\alpha)^{2nd})}\right). \end{aligned} \tag{2.44}$$

2.6.2 Results

Numerical experiments were conducted to investigate the steady-state performance of the greedy algorithm from Section 2.5 relative to three reference schemes: *fixed rate*, *causal genie*, and *noncausal genie*, both with and without finite-buffer constraints at the transmitter. The so-called *fixed-rate* reference scheme chooses the fixed rate (i.e., constellation size) that maximizes expected goodput under the prior SNR distribution, i.e., $\arg\max_{r_t} \int G(r_t, \gamma_t)p(\gamma_t)d\gamma_t$. In the absence of feedback, this fixed rate would be optimal, i.e., total-goodput maximizing. The *causal genie* reference scheme defined in Section 2.4 adapts the rate to maximize expected goodput under perfect causal feedback of the error rate ϵ_t or, equivalently, the SNR γ_t. As shown in Section 2.4, the goodput attained by the causal genie upper bounds that of optimal rate selection under degraded feedback. However, as the feedback delay d and/or the block size n increases, the causal genie's ability to predict the SNR decreases, and thus its goodput suffers. The so-called *non-causal genie* reference scheme assumes perfect knowledge of SNR γ_t for all past, current, and future packets, and uses this information to choose the goodput-maximizing rate. Since this scheme has access to

more information than the causal-genie and greedy algorithms, it upper bounds them in terms of goodput.

Infinite Buffer Experiments

For the first set of experiments, we assumed an infinitely back-logged queue at the transmitter. Unless otherwise noted, the following parameters were used: block size $n = 1$ packet, feedback delay $d = 1$ packet, mean SNR $\mathrm{E}\{\gamma_t\} = 25$ dB, and fading-rate parameter $\alpha = 0.001$. For each channel realization, 200 packets (each consisting of $p = 100$ symbols) were transmitted. The steady-state goodputs reported (per symbol per packet) in the figures were calculated by averaging instantaneous goodputs over the packets in 1000 channel realizations for Figs. 2.3-2.4 and 500 channel realizations for Figs. 2.5-2.6. To ensure that steady-state performance was reported, the algorithms were initialized at the goodput-maximizing rate for each new channel realization.

Figure 2.3 plots steady-state goodput as a function of mean SNR $\mathrm{E}\{\gamma_t\}$. To vary $\mathrm{E}\{\gamma_t\}$, we varied the parameter K while keeping $\alpha = 0.01$. The plot shows the greedy algorithm exhibits an increasing gain over the fixed-rate algorithm as mean SNR increases. At low mean SNR, little gain is observed because the optimal constellation size is almost always the smallest one, as can be inferred from Fig. 2.2. But, at higher mean SNRs, the greedy algorithm performs about 1 dB worse (in SNR) than the causal genie, whereas the fixed-rate scheme performs about 5 dB worse. Furthermore, the SNR gap between the greedy and fixed-rate schemes grows as mean SNR increases. Since the steady goodput achieved by the causal genie upper bounds that achievable by any causal-feedback-based rate adaptation algorithm, one can infer that greedy adaptation based on 1-bit ACK/NAK feedback is sufficient to attain a major fraction of the gain achievable by any causal feedback scheme.

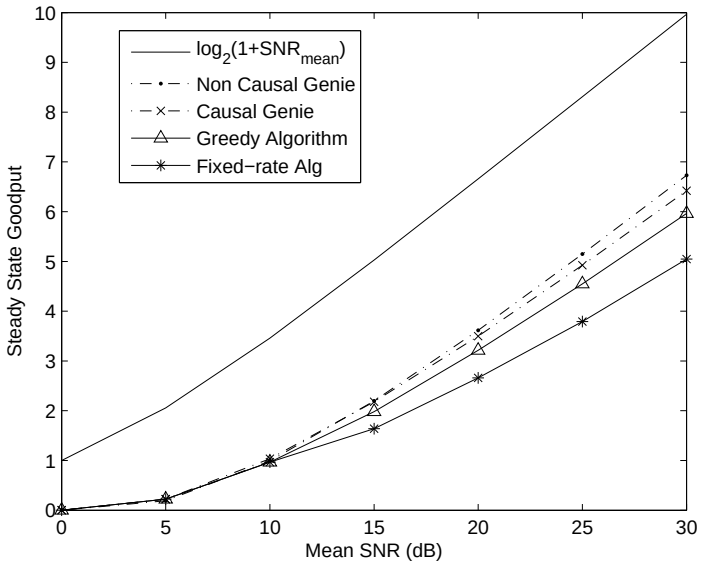

Figure 2.3: Steady-state goodput versus mean SNR $\mathrm{E}\{\gamma_t\}$ for $\alpha = 0.01$, block size $n = 1$ packet, and delay $d = 1$ packet.

Figure 2.4 shows steady-state goodput versus fading-rate parameter α for mean SNR $\mathrm{E}\{\gamma_t\} = 25$ dB. Lower α corresponds to slower channel variation and thus more accurate prediction of instantaneous SNR. From the plot, the following can be observed: as α decreases, both the causal genie and the greedy algorithm approach the non-causal genie, whereas as α increases, both the causal genie and the greedy algorithm approach the fixed-rate algorithm. The non-causal genie and fixed-rate algorithms yield essentially constant[9] steady-state goodput versus α. For a wide range of α, it can be seen that the greedy algorithm performs closer to the causal genie than it does to the fixed-rate algorithm. Thus, we conclude that the greedy scheme captures a dominant fraction (e.g., $\approx 90\%$ at low α) of the goodput gain achievable under causal feedback.

[9] Deviations from constant are due to finite averaging effects.

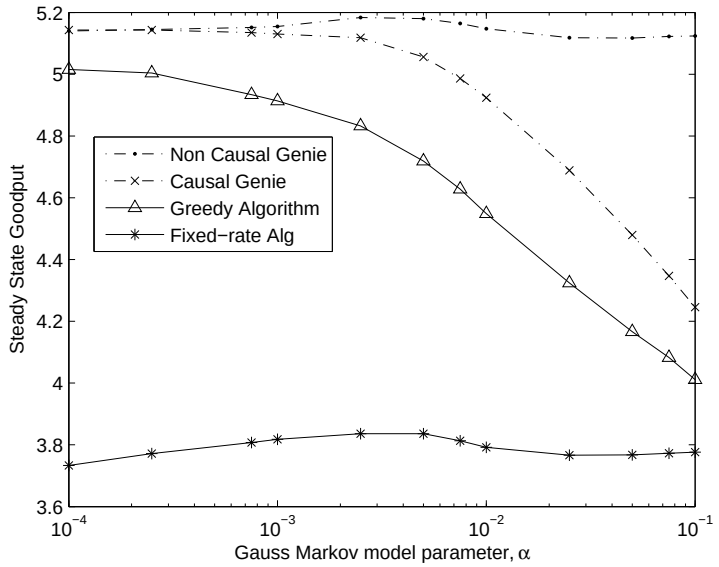

Figure 2.4: Steady-state goodput versus α for $\mathrm{E}\{\gamma_t\} = 25$ dB, block size $n = 1$ packet, and delay $d = 1$ packet.

Figure 2.5 plots steady-state goodput versus feedback delay d for packet-rate adaptation, i.e., $n = 1$. By definition, the non-causal genie has access to all past, current, and future SNRs, so its performance is unaffected by delay. As for the causal genie and greedy algorithms, their steady-state goodputs measure 30% and 20% above that of the fixed-rate algorithm, respectively, when $d = 1$. However, as the delay d increases, their causally predicted SNR distributions converge to the prior SNR distribution, so that, the causal genie and greedy algorithms eventually perform no better than the fixed-rate algorithm. Still, for all delays, the simple greedy scheme captures a dominant fraction of the goodput gain achievable under causal feedback.

Figure 2.6 plots steady-state goodput versus block size n packets for delay $d = 1$ packet and $\alpha = 0.001$. For all tested block sizes, the greedy algorithm performs closer to the causal genie than to the fixed-rate algorithm, implying that the greedy

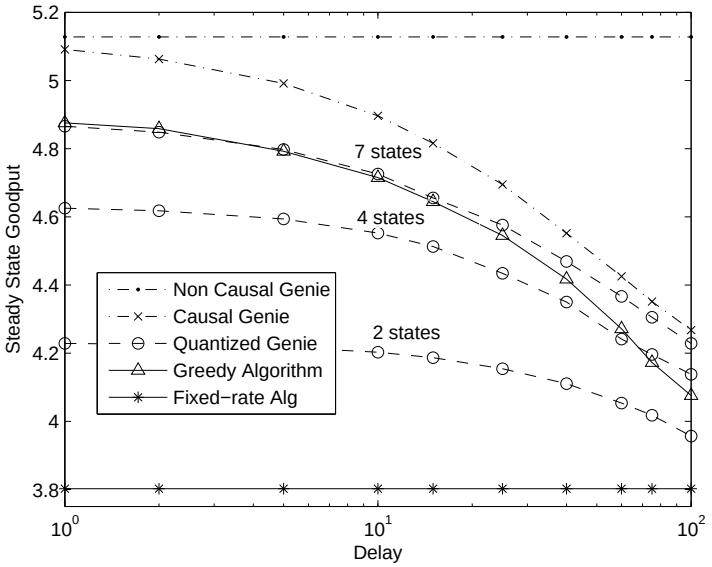

Figure 2.5: Steady-state goodput versus delay d for $\mathrm{E}\{\gamma_t\} = 25$ dB, $\alpha = 0.001$, and block size $n = 1$ packet.

algorithm once again recovers a dominant portion of the goodput gain achievable under the causal feedback constraint. The performances of all adaptive schemes decrease with block size, though. This is for two reasons: first, a uniform rate is applied across the block, whereas the optimal rate varies across the block; and, second, as the block length increases, the SNR must be predicted farther into the future. Notice that even the performance of non-causal genie degrades as n increases due to the sub-optimality of its uniform rate assignment across the block.

Figures 2.5 and 2.6 also plot the performance of the so-called *quantized genie* reference scheme, which adapts the rate to maximize goodput under quantized, but otherwise perfect, knowledge of SNR γ_{t-d}. The goodput attained by the quantized

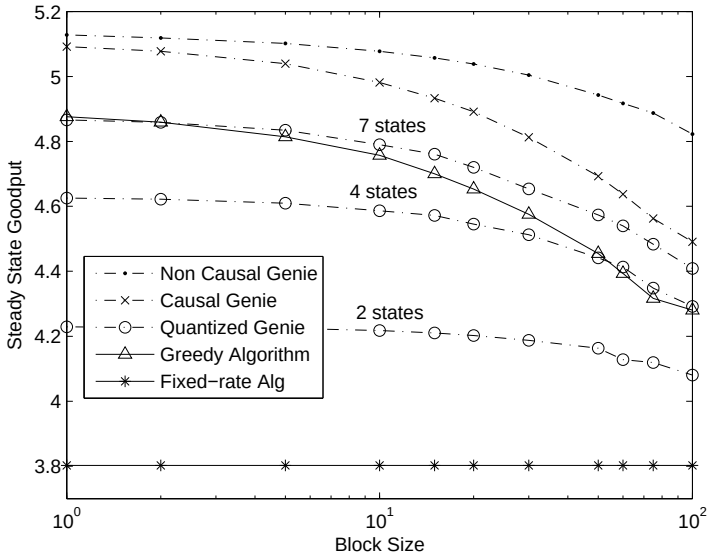

Figure 2.6: Steady-state goodput versus block size n for $\mathrm{E}\{\gamma_t\} = 25$ dB, $\alpha = 0.001$, and delay $d = 1$ packet.

genie upper bounds[10] the goodput attained by a transmitter that assumes a *finite-state* Markov SNR model and employs optimal POMDP-based rate assignment. To construct the corresponding finite-state Markov model, we quantized the SNR using the Lloyd-Max algorithm [23] and calculated the state transition probability matrix via [24, eq. (15)-(16)]. Apart from the finite-state SNR model, rate assignment for the quantized genie is identical to that for the causal genie.

Figures 2.5 and 2.6 show that the greedy algorithm outperforms the 2- and 4-state quantized genies, and performs on par with the 7-state quantized genie, throughout most of the examined range of d and n. Thus, we conclude that the greedy algorithm outperforms the optimal POMDP-based rate adaptation scheme based on a

[10] The fact that the quantized genie yields an upper bound in the case of a finite-state Markov channel follows directly from Lemma 1, which holds for both continuous and finite-state Markov channels.

finite-state Markov SNR model with 7 states or less. This is notable because the computational complexity of optimal POMDP-based rate adaptation is significant, under typical horizons, for channel models with more than a few states.

Finite Buffer Experiments

For this second set of experiments, a finite data buffer was employed at the transmitter. Bits are removed from the buffer when an ACK arrives, confirming their successful transmission, or when the buffer overflows. The following parameters were used: block size $n = 1$ packet, feedback delay $d = 1$ packet, and mean SNR $\mathrm{E}\{\gamma_t\} = 25$ dB. The packet arrival rate followed a 2-state Markov model with ON and OFF states. In the ON state, a single packet arrives in the buffer (queue), and in the OFF state, no packets arrive. The self transition probability in both ON and OFF states was set to 0.9 in order to mimic bursty traffic. Consequently, the steady-state probability of each state is 0.5 and the long-term arrival rate is 0.5 packets/interval. The size of an arriving packet was set equal to the number of bits transmitted (per packet interval) by the fixed-rate reference scheme under backlogged conditions. The size of the buffer was set equal to 30 such packets of data. Thus, if packets were arriving persistently, then, in the absence of NAKs, the fixed-rate scheme would yield a fixed buffer occupancy, while, in the absence of ACKs, the buffer would go from totally empty to totally full after 30 arrivals. For each channel realization, 1000 packets were transmitted (each consisting of $p = 100$ symbols) and the buffer was initialized at half-full. The values reported in the figures represent the average of all packets in 1000 channel realizations.

Figure 2.7 plots average buffer occupancy versus fading-rate parameter α, where a buffer occupancy of "b" is to be interpreted as b arrival-packets worth of bits. It

can be seen that the buffer occupancy achieved by the greedy algorithm is very close to that achieved by the causal and non-causal genie algorithms, whereas the buffer occupancy achieved by the fixed-rate scheme is much higher, especially at lower values of α. Recall that, when α is low, the SNR can remain below average for prolonged periods of time, during which fixed-rate transmissions are more likely to yield NAKs and hence fill the buffer. Figure 2.8 plots a related statistic: the fraction of packets that are dropped due to buffer overflows. Here again, the drop rate achieved by the greedy algorithm is very close to that achieved by the causal and non-causal genie algorithms, whereas the drop rate achieved by the fixed-rate algorithm is more than 10 times higher.

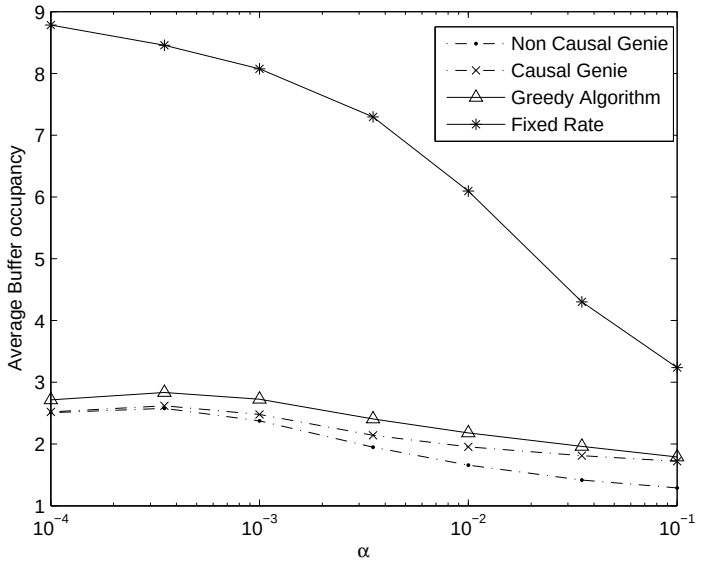

Figure 2.7: Average buffer occupancy versus α for Markov arrivals with average rate = 0.5 packets/interval, buffer size = 30 packets, $\mathrm{E}\{\gamma_t\}$ = 25 dB, block size $n = 1$ packet, and delay $d = 1$ packet.

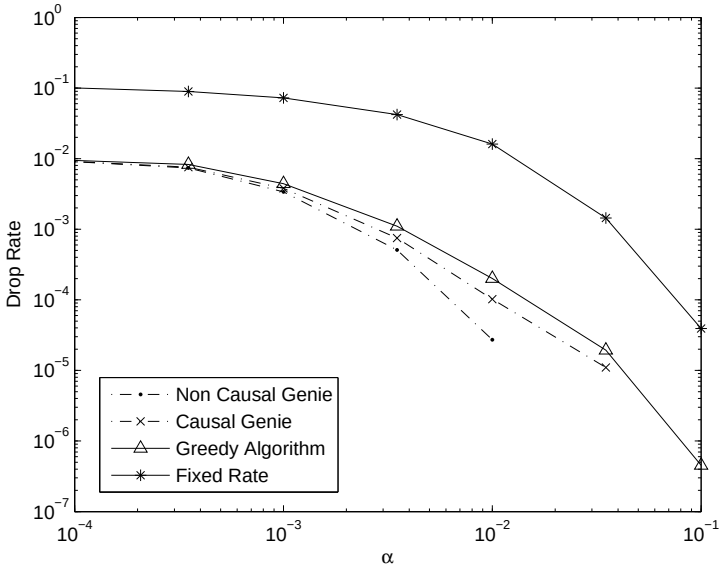

Figure 2.8: Average drop rate versus α for Markov arrivals with average rate = 0.5 packets/interval, buffer size = 30 packets, $\mathrm{E}\{\gamma_t\} = 25$ dB, block size $n = 1$ packet, and delay $d = 1$ packet.

Figure 2.9 shows steady-state goodput versus fading-rate parameter α for Markov arrivals and finite buffer size. The steady-state goodput achieved by the greedy scheme is very close to that of the causal and non-causal genie schemes, whereas the steady-state goodput achieved by the fixed-rate scheme is much lower, especially when α is small. The increase of steady-state goodput with α is directly related to the decrease in drop rate with α observed in Fig. 2.8, since dropped packets do not contribute to goodput.

2.7 Summary

In this chapter, we studied rate adaptation schemes that use degraded error-rate feedback (e.g., packet-rate ACK/NAKs) to maximize finite-horizon expected goodput

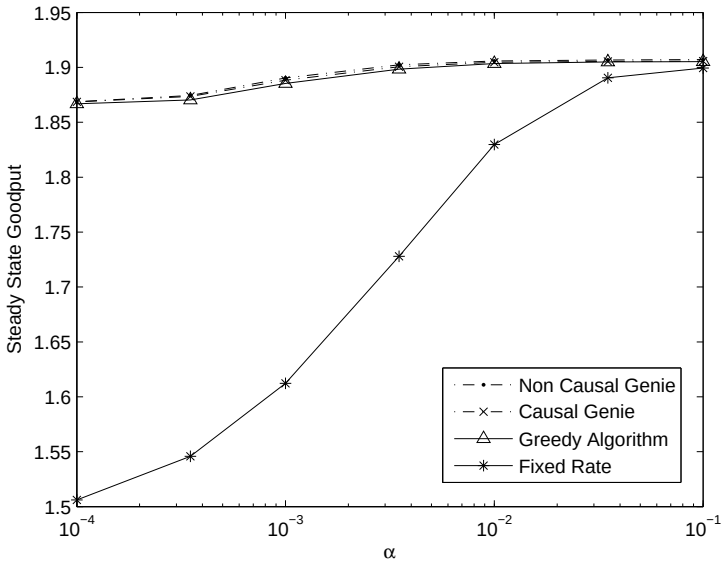

Figure 2.9: Steady-state goodput rate versus α for Markov arrivals with average rate = 0.5 packets/interval, buffer size = 30 packets, $\mathrm{E}\{\gamma_t\}$ = 25 dB, block size $n = 1$ packet, and delay $d = 1$ packet.

over continuous Markov flat-fading wireless channels. First, we specified the POMDP that leads to the optimal rate schedule and showed that its solution is computationally impractical. Then, we proposed a simple greedy alternative and showed that, while generally suboptimal, the greedy approach is optimal when the error-rate feedback is non-degraded. We then detailed an implementation of the greedy rate-adaptation scheme in which the SNR distribution is estimated online (from degraded error-rate feedback) and combined with offline-calculated goodput-versus-SNR curves to find the expected-goodput maximizing transmission rate. In addition to the packet-rate greedy adaptation scheme, a block-rate greedy adaptation scheme was also proposed that offers the potential for significant reduction in complexity with only moderate sacrifice in performance.

For the particular case of uncoded square-QAM transmission, packet-rate ACK/NAK feedback, and Rayleigh fading, the greedy scheme was numerically compared to three reference schemes: the optimal fixed-rate scheme, a genie-aided scheme with perfect causal SNR knowledge, and a genie-aided scheme with perfect non-causal SNR knowledge. First, the effects of mean SNR, channel fading rate, and feedback delay on steady-state goodput were investigated in the context of an infinitely backlogged transmission queue. In this case, the causal genie reference is especially meaningful because it upper bounds the performance of the optimal POMDP scheme, which is too complex to implement directly. Second, a finite transmission buffer was considered, and the effects of channel fading rate on buffer occupancy, drop rate, and steady-state goodput were investigated. The results suggest that the simple packet-rate greedy scheme captures a dominant fraction of the achievable goodput under causal feedback, whereas the optimal fixed-rate scheme captures significantly less. Similarly, the drop rate and average buffer occupancy of the greedy scheme were nearly equal to those of the causal and non-causal genie-aided schemes, whereas the drop rate and average buffer occupancy of the fixed-rate scheme were much higher (e.g., an order-of-magnitude higher in the case of drop rate). Comparisons to a "quantized genie" scheme that upper bounds optimal adaptation under a finite-state Markov SNR model were also made, and there it was found that the proposed greedy scheme outperformed the quantized genie scheme with up to 7 states. Since POMDP-based optimal rate-adaptation for discrete-Markov channels with 7 or more states would be computationally intensive, greedy rate-adaptation based on a continous-Markov channel model is more appealing.

Chapter 3: Joint Scheduling and Resource Allocation in the OFDMA Downlink under Imperfect Channel-State Information

3.1 Introduction

In the downlink of a wireless orthogonal frequency division multiple access (OFDMA) system, the base station (BS) delivers data to a pool of users whose channels vary in both time and frequency. Since bandwidth and power resources are limited, the BS would like to allocate them most effectively, e.g., by pairing users with strong subchannels and distributing power in the best possible manner. At the same time, the BS may need to maintain per-user quality-of-service (QoS) constraints, such as a minimum reliable rate for each user. Overall, the BS faces a resource allocation problem where the goal is to maximize an efficiency-related quantity (e.g., a function of goodput) under particular (e.g., power) constraints [25]. Although, for resource allocation, one would ideally like to have access to instantaneous channel state information (CSI), such CSI is difficult to obtain in practice, and so resource allocation must be accomplished under imperfect CSI. Thus, in this chapter, we consider simultaneous user-scheduling, power-allocation, and rate-selection in an OFDMA downlink, given only a generic *distribution* for the subchannel signal-to-noise ratios (SNRs), with the goal of maximizing expected sum-utility under a sum-power constraint. In doing so,

we consider relatively generic goodput-based utilities, facilitating, e.g., throughput-based pricing (e.g., [26–28]), quality-of-service enforcement, and/or the treatment of practical modulation-and-coding schemes (MCS).

In particular, we consider the above scheduling and resource allocation (SRA) problem under two scenarios. In the first scenario, we allow multiple users (and/or MCSs) to time-share any given subchannel and time-slot. In practice, this scenario occurs, e.g., in OFDMA systems where several users are multiplexed within a time-slot, such as IEEE 802.16/WiMAX [29] and 3GPP LTE [30]. Although the resulting optimization problem is non-convex, we show that it can be converted into a convex problem and solved exactly using a dual optimization approach. Based on a detailed analysis of the optimal solution, we propose a novel bisection-based algorithm that is faster than state-of-the-art golden-section based approaches (e.g., [31]) and that admits finite-iteration performance guarantees. In the second scenario, we allow at most one combination of user and MCS to be used on any given subchannel and time-slot. This scenario occurs widely in practice, such as in the Dedicated Traffic Channel (DTCH) mode of UMTS-LTE [32], and results in a mixed-integer optimization problem. Based on a detailed analysis of the optimal solution to this problem and its relationship to that in the first scenario, we propose a novel sub-optimal algorithm that is faster than state-of-the-art golden-section and subgradient based approaches (e.g., [31, 33]), and we derive a novel tight bound on the optimality gap of our algorithm. Finally, we simulate our algorithms under various OFDMA system configurations, comparing against state-of-the-art approaches and genie-aided performance bounds.

The remainder of this chapter is organized as follows. In Section 3.2, we discuss some past work that is related to our problem. In Section 3.3, we outline the system model and frame our optimization problems. In Section 3.4, we consider the "continuous" problem, where each subchannel can be shared by multiple users and rates, and find its exact solution. In Section 3.5, we consider the "discrete" problem, where each subchannel can support at most one combination of user and rate per time slot. In Section 3.6, we compare the performance of the proposed algorithms to reference algorithms under various settings. Finally, in Section 3.7, we conclude.

3.2 Past Work

The problem of OFDMA downlink SRA under *perfect* CSI has been studied in several papers, notably [34–39]. In [34], a utility maximization framework for discrete allocation was formulated to balance system efficiency and fairness, and efficient subgradient-based algorithms were proposed. In [35], a subchannel, rate, and power allocation algorithm was developed to minimize power consumption while maintaining a total rate-allocation requirement for every user. In [38], a weighted-sum capacity maximization problem with/without subchannel sharing was formulated to allocate subcarriers and powers. In [39], non-convex optimization problems regarding weighted sum-rate maximization and weighted sum-power minimization were solved using a Lagrange dual decomposition method. Compared to the above works, we extend the utility maximization framework to imperfect CSI and continuous allocations, and propose bisection-based algorithms that are faster for both the discrete and continuous allocation scenarios. Unlike [34–39], our utility framework can be

applied to problems with/without fixed rate-power functions[11]. In additional, it can be applied to pricing-based utilities (e.g., responsive pricing and proportional fairness pricing) [26]. Furthermore, we study the relationship between the discrete and continuous allocation scenarios, and provide a tight bound on the duality gap of our proposed discrete-allocation scheme.

The problem of OFDMA downlink SRA under *imperfect* CSI was studied in several papers, notably [31,33,40,41]. In [33], the authors considered the problem of discrete ergodic weighted sum-rate maximization for user scheduling and resource allocation, and studied the impact of channel estimation error due to pilot-aided MMSE channel estimation. In [31], a deterministic optimization problem was formulated using an upper bound on system capacity (via Jensen's inequality) as the objective. Both optimal and heuristic algorithms were then proposed to implement the obtained solution. Compared to these two works, we propose faster algorithms, applicable to a general utility maximization framework (of which the objectives in [31,33] are special cases), under a more general class of channel estimators, and for both discrete and continuous subchannel allocations. Our algorithms are inspired by a rigorous analysis of the optimal solutions to the discrete and continuous problems. In [40], the problem of total transmit power minimization, subject to strict constraints on conditional expected user capacities, was investigated. In [41], the effect of heterogeneous delay requirements and outdated CSI on a particular discrete resource allocation problem was studied. In contrast, we consider a general utility maximization problem that allows us to attack problems that may or may not be based on fixed rate-power functions, as well as those based on pricing models. Relative to these works, we propose

[11]By a "fixed rate-power function" we mean that, for a given SNR, the achievable rate is a known function of the power.

faster algorithms for both continuous and discrete allocation problems with provable bounds on their performances.

3.3 System Model

Consider a downlink OFDMA system with N subchannels and K active users ($N, K \in \mathbb{Z}^+$) as shown in Fig. 3.1. The scheduler-and-resource-allocator at the base-station uses the imperfect CSI to send data to the users, across OFDMA subchannels, in a way that maximizes utility. We assume that, for each user, there is an infinite backlog of data at the base-station, so that there is always data available to be transmitted. During every channel use and across every OFDMA subchannel, the base-station transmits codeword(s) from a generic signaling scheme, which propagate to the intended mobile recipient(s) through their respective fading channels. For a given user k, the OFDMA subchannels are assumed to be non-interfering, with gains that are time-invariant over each codeword duration and statistically independent of those for other users. Thus, the successful reception of a transmitted codeword depends on the corresponding subchannel's SNR γ, power p, and modulation and coding scheme (MCS), indexed by $m \in \{1, \ldots, M\}$. We assume that, for user k, MCS m corresponds to a transmission rate of $r_{k,m}$ bits per codeword and a codeword error probability of $\epsilon_{k,m}(p\gamma) = a_{k,m}e^{-b_{k,m}p\gamma}$ for known constants $a_{k,m}$ and $b_{k,m}$ (see, e.g., [33]). Here, the subchannel SNR γ is treated as an exogenous parameter, so that $p\gamma$ is the effective received SNR.

To precisely state our scheduling and resource allocation (SRA) problem, some additional notation is useful. To indicate how subchannels are partitioned among users and rates in each time-slot, we will use the proportionality indicator $I_{n,k,m}$,

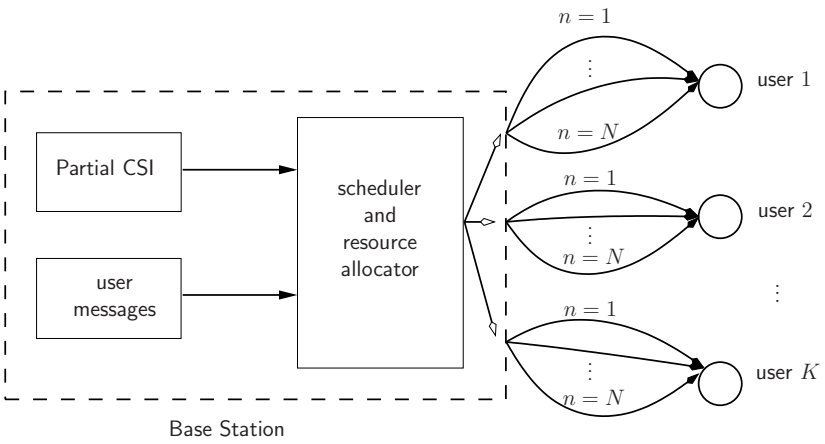

Figure 3.1: System model of a downlink OFDMA system with N subchannels and K users. Here, n is the subchannel index.

where $I_{n,k,m} = 1$ means that subchannel n is fully dedicated to user k at MCS m, and $I_{n,k,m} = 0$ means that subchannel n is totally unavailable to user k at MCS m. The subchannel resource constraint is then expressed as $\sum_{k,m} I_{n,k,m} \leq 1$ for all n. In the sequel, we consider two flavors of the SRA problem, a "continuous" one where each subchannel can be shared among multiple users and/or rates per time slot (i.e., $I_{n,k,m} \in [0, 1]$), and a "discrete" one where each subchannel can be allocated to at most one user/rate combination per time slot (i.e., $I_{n,k,m} \in \{0, 1\}$). We will use $p_{n,k,m} \geq 0$ as the power that would be expended on subchannel n if it was fully allocated to the user/rate combination (k, m). With this definition, the total expended power becomes $\sum_{n,k,m} I_{n,k,m} p_{n,k,m}$. Finally, we will use $\gamma_{n,k}$ to denote the n^{th} subchannel's SNR for user k. Although we assume that the BS does not know the SNR realizations $\{\gamma_{n,k}\}$, we assume that it does know the (marginal) distribution of each $\gamma_{n,k}$.

When subchannel n is fully dedicated to user k with MCS m and power $p_{n,k,m}$, the goodput $g_{n,k,m} = (1 - a_{k,m} e^{-b_{k,m} p_{n,k,m} \gamma_{n,k}}) r_{k,m}$ quantifies the expected number of

bits, per codeword, transmitted without error. In the sequel, we focus on maximizing goodput-based utilities of the form $U_{n,k,m}(g_{n,k,m})$, where $U_{n,k,m}(\cdot)$ is any generic real-valued function that is twice differentiable, strictly-increasing, and concave, with $U_{n,k,m}(0) < \infty$. (These conditions imply $U'_{n,k,m}(\cdot) > 0$ and $U''_{n,k,m}(\cdot) \leq 0$.) In particular, we aim to maximize the expected sum utility, $\mathrm{E}\big\{\sum_{n,k,m} I_{n,k,m} U_{n,k,m}(g_{n,k,m})\big\}$, where the expectation is taken over the subchannel-SNRs $\{\gamma_{n,k}\}$ hidden within the goodputs. Incorporating a sum-power constraint of P_{con}, our SRA problem becomes

$$\begin{aligned}\mathrm{SRA} \quad \triangleq \quad & \max_{\substack{\{p_{n,k,m}\geq 0\}\\ \{I_{n,k,m}\}}} \mathrm{E}\left\{\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M} I_{n,k,m} U_{n,k,m}\big((1-a_{k,m}e^{-b_{k,m}p_{n,k,m}\gamma_{n,k}})r_{k,m}\big)\right\} \\ & \text{s.t.} \quad \sum_{k,m} I_{n,k,m} \leq 1 \ \forall n \quad \text{and} \quad \sum_{n,k,m} I_{n,k,m} p_{n,k,m} \leq P_{\mathsf{con}}. \end{aligned} \tag{3.1}$$

The above formulation is sufficiently general to address a wide class of objectives. For example, to maximize sum-goodput, one would simply use $U_{n,k,m}(g) = g$. For weighted sum-goodput, one would instead choose $U_{n,k,m}(g) = w_k g$ with appropriately chosen weights $\{w_k\}$. To maximize weighted sum capacity $\sum_{n,k} w_k I_{n,k,1} \log\big(1 + p_{n,k,1}\gamma_{n,k}\big)$, as in [33], one would choose $M = a_{k,1} = b_{k,1} = r_{k,1} = 1$, and set $U_{n,k,1}(g) = w_k \log\big(1 - \log(1-g)\big)$ for $g \in [0,1)$. Commonly used utilities constructed from concave functions of capacity $\log(1 + p_{n,k,1}\gamma_{n,k})$, such as max-min fairness and the utilities in [34] and [31], can also be handled by our formulation. For example, the utility $U_{n,k,m}(g) = 1 - e^{-w_k g}$ (for some positive $\{w_k\}$) is appropriate for "elastic" applications such as file transfer [27, 28]. Our formulation also supports various pricing models [26], such as flat-pricing, responsive pricing, proportional fairness pricing, and effective-bandwidth pricing.

Next, in Section 3.4, we study the SRA problem for the continuous case $I_{n,k,m} \in [0,1]$, and in Section 3.5 we study it for the discrete case $I_{n,k,m} \in \{0,1\}$.

3.4 Optimal Scheduling and Resource Allocation with subchannel sharing

In this section, we address the SRA problem in the case where $I_{n,k,m} \in [0,1]\ \forall (n,k,m)$. Recall that this problem arises when sharing of any subchannel by multiple users and/or multiple MCS combinations is allowed. We refer to this problem as the "continuous scheduling and resource allocation" (CSRA) problem. Defining $\mathbf{I}$ as the $N \times K \times M$ matrix with $(n,k,m)^{\text{th}}$ element as $I_{n,k,m}$ and the domain of $\mathbf{I}$ as

$$\mathcal{I}_{\mathsf{CSRA}} := \big\{\mathbf{I} : \mathbf{I} \in [0,1]^{N\times K\times M},\ \textstyle\sum_{k,m} I_{n,k,m} \le 1\ \forall n\big\},$$

the CSRA problem can be stated as

$$\begin{aligned} \text{CSRA} \ := \min_{\substack{\{p_{n,k,m}\ge 0\}\\ \mathbf{I}\in\mathcal{I}_{\mathsf{CSRA}}}} & -\sum_{n,k,m} I_{n,k,m}\, \mathrm{E}\left\{U_{n,k,m}\big((1-a_{k,m}e^{-b_{k,m}p_{n,k,m}\gamma_{n,k}})r_{k,m}\big)\right\} \\ \text{s.t.} & \sum_{n,k,m} I_{n,k,m}\, p_{n,k,m} \le P_{\mathsf{con}}. \end{aligned} \tag{3.2}$$

This problem has a non-convex constraint set, making it a non-convex optimization problem. In order to convert it into a convex optimization problem, we write the "actual" power allocated to user k at MCS m on subchannel n as $x_{n,k,m} = I_{n,k,m}\, p_{n,k,m}$. Then, the problem becomes

$$\text{CSRA} \ = \min_{\substack{\{x_{n,k,m}\ge 0\}\\ \mathbf{I}\in\mathcal{I}_{\mathsf{CSRA}}}} \sum_{n,k,m} I_{n,k,m}\, F_{n,k,m}(I_{n,k,m}, x_{n,k,m}) \quad \text{s.t.} \ \sum_{n,k,m} x_{n,k,m} \le P_{\mathsf{con}}, \tag{3.3}$$

where $F_{n,k,m}(\cdot,\cdot)$ is given by

$$F_{n,k,m}(I_{n,k,m}, x_{n,k,m}) = \begin{cases} -\,\mathrm{E}\left\{U_{n,k,m}\big((1-a_{k,m}e^{-b_{k,m}x_{n,k,m}\gamma_{n,k}/I_{n,k,m}})r_{k,m}\big)\right\} & \text{if } I_{n,k,m}\neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

The modified problem in (3.3) is a convex optimization problem with a convex objective function and linear inequality constraint. Moreover, Slater's condition is satisfied

at $I_{n,k,m} = \frac{1}{2KM}$ and $x_{n,k,m} = \frac{P_{\mathsf{con}}}{N} I_{n,k,m},\ \forall n, k, m$. Hence, the solution of (3.3) is the same as that of its dual problem (i.e., zero duality gap) [42]. Let us denote the optimal $\mathbf{I}$ and $\mathbf{x}$ for (3.3) by $\mathbf{I}^*_{\mathsf{CSRA}}$ and $\mathbf{x}^*_{\mathsf{CSRA}}$, respectively, and let $\mathbf{p}^*_{\mathsf{CSRA}}$ be the corresponding $\mathbf{p}$.

Writing the dual formulation, using μ as the dual variable, the Lagrangian of (3.3) is

$$L(\mu, \mathbf{I}, \mathbf{x}) = \sum_{n,k,m} I_{n,k,m}\, F_{n,k,m}(I_{n,k,m}, x_{n,k,m}) + \Big(\sum_{n,k,m} x_{n,k,m} - P_{\mathsf{con}} \Big)\mu, \tag{3.5}$$

where we use $\mathbf{x}$ to denote the $N \times K \times M$ matrix $[x_{n,k,m}]$. The corresponding unconstrained dual problem, then, becomes

$$\begin{aligned} \max_{\mu \geq 0} \min_{\substack{\mathbf{x} \succeq 0 \\ \mathbf{I} \in \mathcal{I}_{\mathsf{CSRA}}}} L(\mu, \mathbf{I}, \mathbf{x}) &= \max_{\mu \geq 0} \min_{\mathbf{I} \in \mathcal{I}_{\mathsf{CSRA}}} L(\mu, \mathbf{I}, \mathbf{x}^*(\mu, \mathbf{I})) \\ &= \max_{\mu \geq 0} L(\mu, \mathbf{I}^*(\mu), \mathbf{x}^*(\mu, \mathbf{I}^*(\mu))) \\ &= L(\mu^*, \mathbf{I}^*(\mu^*), \mathbf{x}^*(\mu^*, \mathbf{I}^*(\mu^*))), \end{aligned} \tag{3.6}$$

where $\mathbf{x} \succeq 0$ means that $x_{n,k,m} \geq 0\ \forall n, k, m$, $\mathbf{x}^*(\mu, \mathbf{I})$ denotes the optimal $\mathbf{x}$ for a given μ and $\mathbf{I}$, $\mathbf{I}^*(\mu) \in \mathcal{I}_{\mathsf{CSRA}}$ denotes the optimal $\mathbf{I}$ for a given μ, and μ^* denotes the optimal μ.

In the next few subsections, we will optimize the Lagrangian according to (3.6) w.r.t. $\mathbf{x}$, $\mathbf{I}$, and μ in Section 3.4.1, Section 3.4.2, and Section 3.4.3, respectively. We then propose an iterative algorithm to solve CSRA problem in Section 3.4.4. Finally, we discuss some important properties of the CSRA solution in Section 3.4.5.

3.4.1 Optimizing over total powers, x, for a given μ and user-MCS allocation matrix I

The Lagrangian in (3.5) is a convex function of $\mathbf{x}$. Therefore, any local minimum of the function is a global minimum. Calculating the derivative of $L(\mu, \mathbf{I}, \mathbf{x})$ w.r.t.

$x_{n,k,m}$, we get

$$\frac{\partial L(\mu,\mathbf{I},\mathbf{x})}{\partial x_{n,k,m}} \tag{3.7}$$
$$= \begin{cases} \mu & \text{if } I_{n,k,m}=0 \\ \mu - a_{k,m}b_{k,m}r_{k,m}\,\mathrm{E}\Big\{U'_{n,k,m}\big((1-a_{k,m}e^{-b_{k,m}x_{n,k,m}\gamma_{n,k}/I_{n,k,m}}) \times r_{k,m}\big)\gamma_{n,k}e^{-b_{k,m}x_{n,k,m}\gamma_{n,k}/I_{n,k,m}}\Big\} & \text{otherwise.} \end{cases}$$

Clearly, if $I_{n,k,m}=0$, then $L(\cdot,\cdot,\cdot)$ is an increasing[12] function of $x_{n,k,m}$ since $\mu \geq 0$. Therefore, $x^*_{n,k,m}(\mu,\mathbf{I})=0$. But if $I_{n,k,m}\neq 0$, then $\frac{\partial L(\mu,\mathbf{I},\mathbf{x})}{\partial x_{n,k,m}}$ is an increasing function of $x_{n,k,m}$ since $U'_{n,k,m}(\cdot)$ is a decreasing function of $x_{n,k,m}$. Thus, we have

$$\mu - a_{k,m}b_{k,m}r_{k,m}\,\mathrm{E}\Big\{U'_{n,k,m}\big((1-a_{k,m}e^{-b_{k,m}x_{n,k,m}\gamma_{n,k}/I_{n,k,m}})r_{k,m}\big) \times\gamma_{n,k}e^{-b_{k,m}x_{n,k,m}\gamma_{n,k}/I_{n,k,m}}\Big\}=0 \tag{3.8}$$

for some positive $x_{n,k,m}$ if and only if $0\leq\mu\leq a_{k,m}b_{k,m}r_{k,m}U'_{n,k,m}\big((1-a_{k,m})r_{k,m}\big)\,\mathrm{E}\{\gamma_{n,k}\}$. Therefore,

$$x^*_{n,k,m}(\mu,\mathbf{I})=\begin{cases}\tilde{x}_{n,k,m}(\mu,\mathbf{I}) & \text{if } 0\leq\mu\leq a_{k,m}b_{k,m}r_{k,m}U'_{n,k,m}\big((1-a_{k,m})r_{k,m}\big)\,\mathrm{E}\{\gamma_{n,k}\}\\ 0 & \text{otherwise,}\end{cases} \tag{3.9}$$

where $\tilde{x}_{n,k,m}(\mu,\mathbf{I})$ satisfies

$$\mu = a_{k,m}b_{k,m}r_{k,m}\,\mathrm{E}\Big\{U'_{n,k,m}\big((1-a_{k,m}e^{-b_{k,m}\tilde{x}_{n,k,m}(\mu,\mathbf{I})\gamma_{n,k}/I_{n,k,m}})r_{k,m}\big) \times\gamma_{n,k}e^{-b_{k,m}\tilde{x}_{n,k,m}(\mu,\mathbf{I})\gamma_{n,k}/I_{n,k,m}}\Big\}. \tag{3.10}$$

From (3.10), we observe that $\tilde{x}_{n,k,m}(\mu,\mathbf{I})=\tilde{p}_{n,k,m}(\mu)I_{n,k,m}$, where $\tilde{p}_{n,k,m}(\mu)$ satisfies

$$\mu = a_{k,m}b_{k,m}r_{k,m}\,\mathrm{E}\Big\{U'_{n,k,m}\big((1-a_{k,m}e^{-b_{k,m}\tilde{p}_{n,k,m}(\mu)\gamma_{n,k}})r_{k,m}\big) \times\gamma_{n,k}e^{-b_{k,m}\tilde{p}_{n,k,m}(\mu)\gamma_{n,k}}\Big\}. \tag{3.11}$$

[12]We use the terms "increasing" and "decreasing" interchangeably with "non-decreasing" and "non-increasing", respectively. The terms "strictly-increasing" and "strictly-decreasing" are used when appropriate.

Combining the above observations, we can write for any $\mathbf{I} \in \mathcal{I}_{\mathsf{CSRA}}$ and (n,k,m) that

$$x^*_{n,k,m}(\mu, \mathbf{I}) = I_{n,k,m}\, p^*_{n,k,m}(\mu), \tag{3.12}$$

where

$$p^*_{n,k,m}(\mu) = \begin{cases} \tilde{p}_{n,k,m}(\mu) & \text{if } 0 \le \mu \le a_{k,m} b_{k,m} r_{k,m} U'_{n,k,m}\big((1-a_{k,m}) r_{k,m}\big)\, \mathrm{E}\{\gamma_{n,k}\} \\ 0 & \text{otherwise,} \end{cases} \tag{3.13}$$

and $\tilde{p}_{n,k,m}(\mu)$ satisfies (3.11). Note that if such a $\tilde{p}_{n,k,m}(\mu)$ exists that satisfies (3.11), then it is unique. This is because, in (3.11), $U'_{n,k,m}(\cdot)$ is a continuous decreasing positive function and $e^{-b_{k,m}\tilde{p}_{n,k,m}(\mu)\gamma_{n,k}}$ is a strictly-decreasing continuous function of $\tilde{p}_{n,k,m}(\mu)$, which makes the right side of (3.11) a strictly-decreasing continuous function of $\tilde{p}_{n,k,m}(\mu)$. Therefore, in the domain of its existence, $\tilde{p}_{n,k,m}(\mu)$ is unique and decreases continuously with increase in μ. Consequently, $x^*_{n,k,m}(\mu, \mathbf{I})$ is a decreasing continuous function of μ. Figure 3.2 shows an example of the variation of $p^*_{n,k,m}(\mu)$ w.r.t. μ.

3.4.2 Optimizing over user-MCS allocation matrix I for a given μ

Substituting $\mathbf{x}^*(\mu, \mathbf{I})$ from (3.12) into (3.5), we get the Lagrangian

$$\begin{aligned} & L(\mu, \mathbf{I}, \mathbf{x}^*(\mu, \mathbf{I})) \qquad (3.14) \\ &= -\mu P_{\mathsf{con}} + \sum_n \underbrace{\sum_{k,m} I_{n,k,m} \Bigg[\overbrace{-\mathrm{E}\left\{ U_{n,k,m}\big((1 - a_{k,m} e^{-b_{k,m} p^*_{n,k,m}(\mu)\gamma_{n,k}}) r_{k,m}\big) \right\}}^{V_{n,k,m}(\mu, p^*_{n,k,m}(\mu))} + \mu p^*_{n,k,m}(\mu) \Bigg]}_{L_n(\mu, \mathbf{I}_n)}, \end{aligned}$$

where $\mathbf{I}_n = \{I_{n,k,m}\ \forall (k,m)\}$. Since the above Lagrangian contains the sum of $L_n(\mu, \mathbf{I}_n)$ over n, minimizing $L_n(\mu, \mathbf{I}_n)$ for every n (over all possible $\mathbf{I}_n$) minimizes the Lagrangian. Recall that $L_n(\mu, \mathbf{I}_n)$ is a linear function of $\{I_{n,k,m}\ \forall (k,m)\}$ that

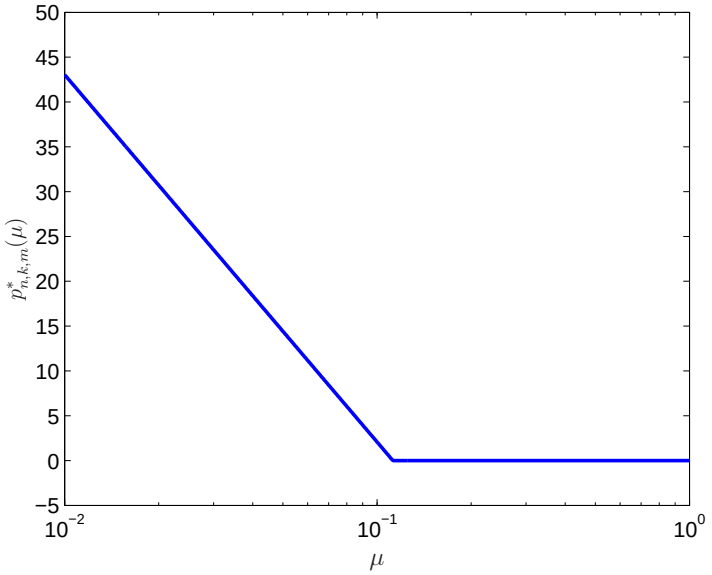

Figure 3.2: Prototypical plot of $p^*_{n,k,m}(\mu)$ as a function of μ. The choice of system parameters are the same as those used in Section 3.6.

satisfies $\sum_{k,m} I_{n,k,m} \leq 1$. Therefore, $L_n(\mu, \mathbf{I}_n)$ is minimized by the $\mathbf{I}_n$ that gives maximum possible weight to the (k,m) combination with the most negative value of $V_{n,k,m}(\mu, p^*_{n,k,m}(\mu))$. To write this mathematically, let us define, for each μ and subchannel n, a set of participating user-MCS combinations that yield the same most-negative value of $V_{n,k,m}(\mu, p^*_{n,k,m}(\mu))$ over all (k,m) as follows:

$$S_n(\mu) \tag{3.15}$$
$$\triangleq \left\{(k,m) : (k,m) = \underset{(k',m')}{\operatorname{argmin}} V_{n,k',m'}(\mu, p^*_{n,k',m'}(\mu)), \text{ and } V_{n,k,m}(\mu, p^*_{n,k,m}(\mu)) \leq 0\right\}.$$

If $S_n(\mu)$ is a null or a singleton set, then the optimal allocation on subchannel n is given by

$$I^*_{n,k,m}(\mu) = \begin{cases} 1 & \text{if } (k,m) \in S_n(\mu) \\ 0 & \text{otherwise.} \end{cases} \tag{3.16}$$

However, if $|S_n(\mu)| > 1$ (where $|S_n(\mu)|$ denotes the cardinality of $S_n(\mu)$), then multiple (k, m) combinations contribute equally towards the minimum value of $L_n(\mu, \mathbf{I})$, and thus the optimum can be reached by sharing subchannel n. In particular, let us suppose that $S_n(\mu) = \{\big(k_1(n), m_1(n)\big), \ldots, \big(k_{|S_n(\mu)|}(n), m_{|S_n(\mu)|}(n)\big)\}$. Then, the optimal allocation of subchannel n is given by

$$I^*_{n,k,m}(\mu) = \begin{cases} I_{n,k_i(n),m_i(n)} & \text{if } (k,m) = (k_i(n), m_i(n)) \text{ for some } i \in \{1, \ldots, |S_n(\mu)|\} \\ 0 & \text{otherwise,} \end{cases} \tag{3.17}$$

where the vector $(I_{n,k_1(n),m_1(n)}, \ldots, I_{n,k_{|S_n(\mu)|}(n),m_{|S_n(\mu)|}(n)})$ is any point in the unit-$(|S_n(\mu)| - 1)$ simplex, i.e., it belongs to the space $[0,1]^{|S_n(\mu)|}$ and satisfies

$$\sum_{i=1}^{|S_n(\mu)|} I_{n,k_i(n),m_i(n)} = 1. \tag{3.18}$$

3.4.3 Optimizing over μ

In order to optimize over μ, we can calculate the Lagrangian optimized for a given value of μ as

$$\begin{aligned} & L(\mu, \mathbf{I}^*(\mu), \mathbf{x}^*(\mu, \mathbf{I}^*(\mu))) \\ &= \sum_{n,k,m} I^*_{n,k,m}(\mu) \Big[- \mathrm{E}\left\{ U_{n,k,m}\big((1 - a_{k,m} e^{-b_{k,m} p^*_{n,k,m}(\mu)\gamma_{n,k}}) r_{k,m}\big)\right\} + \mu p^*_{n,k,m}(\mu) \Big] \\ &\quad - \mu P_{\mathsf{con}}, \end{aligned} \tag{3.19}$$

and then maximize it over all possible values of $\mu \geq 0$ to find μ^*. Notice from (3.16)-(3.18) that we have $\sum_{k,m} I^*_{n,k,m}(\mu^*) = 1$ for at least one n. Otherwise, $\mathbf{I}^*(\mu^*) = \mathbf{0}$ which, clearly, is not the optimal solution. Therefore, $\mu^* \geq \mu_{\mathsf{min}} > 0$, where

$$\mu_{\mathsf{min}} = \min_{n,k,m} a_{k,m} b_{k,m} r_{k,m} \, \mathrm{E}\left\{ U'_{n,k,m}\big((1 - a_{k,m} e^{-b_{k,m} P_{\mathsf{con}} \gamma_{n,k}}) r_{k,m}\big) \gamma_{n,k} e^{-b_{k,m} P_{\mathsf{con}} \gamma_{n,k}} \right\} \tag{3.20}$$

is obtained by taking $\tilde{p}_{n,k,m}(\mu) \to P_{\mathsf{con}}$ for all (n,k,m) in the right side of (3.11). Since $p^*_{n,k,m}(\mu)$ is a decreasing continuous function of μ (seen in Section 3.4.1), we have $\sum_{n,k,m} x^*_{n,k,m}(\mu, \mathbf{I}) > P_{\mathsf{con}}$ for all $\mathbf{I} \neq \mathbf{0}$ and $\mu < \mu_{\mathsf{min}}$. We can also obtain an upper bound $\mu^* \leq \mu_{\mathsf{max}}$, where

$$\mu_{\mathsf{max}} = \max_{n,k,m} a_{k,m} b_{k,m} r_{k,m} U'_{n,k,m}\big((1-a_{k,m}) r_{k,m}\big)\, \mathrm{E}\{\gamma_{n,k}\} \tag{3.21}$$

is obtained by taking $\tilde{p}_{n,k,m}(\mu) \to 0$ in the right side of (3.11). Thus, for any $\mu > \mu_{\mathsf{max}}$, we have that $x^*_{n,k,m}(\mu, \mathbf{I}) = 0 \ \forall n,k,m,\mathbf{I}$. Since the primal objective in (3.3) is certainly not maximized when zero power is allocated on all subchannels, we have $\mu^* \in [\mu_{\mathsf{min}}, \mu_{\mathsf{max}}] \subset (0, \infty)$.

At the optimal μ, i.e., μ^*, if we have $|S_n(\mu^*)| \leq 1 \ \forall n$, then the optimal CSRA allocation, $\mathbf{I}^*_{\mathsf{CSRA}}$, equals $\mathbf{I}^*(\mu^*)$ and can be calculated using (3.16). Moreover, the optimal power allocation $\mathbf{p}^*_{\mathsf{CSRA}}$ allocates

$$p^*_{n,k,m,\mathsf{CSRA}} = \begin{cases} p^*_{n,k,m}(\mu^*) & \text{if } I^*_{n,k,m}(\mu^*) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.22}$$

to every possible (n,k,m) combination. However, if for some n, we have $|S_n(\mu^*)| > 1$, then ambiguity arises due to multiple possibilities of $\mathbf{I}^*(\mu^*)$ obtained via (3.17). In order to find the optimal user-MCS allocation in such cases, we use the fact that the CSRA problem in (3.3) is a convex optimization problem whose exact solution satisfies the sum-power constraint with equality, i.e.,

$$\sum_{n,k,m} x^*_{n,k,m}(\mu^*, \mathbf{I}^*(\mu^*)) = \sum_{n,k,m} I^*_{n,k,m}(\mu^*) p^*_{n,k,m}(\mu^*) = P_{\mathsf{con}}. \tag{3.23}$$

This is because $\mu^* \geq \mu_{\mathsf{min}} > 0$ (shown earlier) and the complementary slackness condition gives that $\mu^*\big(\sum_{n,k,m} x^*_{n,k,m}(\mu^*, \mathbf{I}^*(\mu^*)) - P_{\mathsf{con}}\big) = 0$. Now, recall that the total power allocated to any subchannel n at μ^* is $\sum_{i=1}^{|S_n(\mu^*)|} I_{n,k_i(n),m_i(n)}\, p^*_{n,k_i(n),m_i(n)}(\mu^*)$

where $\{I_{n,k_i(n),m_i(n)}\}_{i=1}^{|S_n(\mu^*)|}$ satisfies (3.18). This quantity is dependent on the choice of values for $\{I_{n,k_i(n),m_i(n)}\}_{i=1}^{|S_n(\mu^*)|}$ and takes on any value between an upper and lower bound given by the following equation:

$$\min_i \, p^*_{n,k_i(n),m_i(n)}(\mu^*) \leq \sum_{i=1}^{|S_n(\mu^*)|} I_{n,k_i(n),m_i(n)} \, p^*_{n,k_i(n),m_i(n)}(\mu^*) \leq \max_i \, p^*_{n,k_i(n),m_i(n)}(\mu^*). \tag{3.24}$$

Note that the existence of at least one $\mathbf{I} = \mathbf{I}^*(\mu^*)$ satisfying

$$\sum_n \sum_i I_{n,k_i(n),m_i(n)} \, p^*_{n,k_i(n),m_i(n)}(\mu^*) = P_{\mathsf{con}} \tag{3.25}$$

is guaranteed by the optimality of the dual solution (of our convex CSRA problem over a closed constraint set). Therefore, we necessarily have $\sum_n \min_i p^*_{n,k_i(n),m_i(n)}(\mu^*) \leq P_{\mathsf{con}}$, and $\sum_n \max_i p^*_{n,k_i(n),m_i(n)}(\mu^*) \geq P_{\mathsf{con}}$. In addition, all choices of user-MCS allocations, $\mathbf{I}^*(\mu^*)$, given by (3.17) that satisfy the equality $\sum_{n,k,m} I^*_{n,k,m}(\mu^*) \, p^*_{n,k,m}(\mu^*) = P_{\mathsf{con}}$, are optimal for the CSRA problem.

In the case that the optimal solution $\mathbf{I}^*(\mu^*)$ is non-unique, i.e., $|S_n(\mu^*)| > 1$ for some n, then one instance of $\mathbf{I}^*(\mu^*)$ can be found as follows. For each subchannel n, define

$$(k_{\mathsf{max}}(n,\mu^*), m_{\mathsf{max}}(n,\mu^*)) \;:=\; \operatorname*{argmax}_i \, p^*_{n,k_i(n),m_i(n)}(\mu^*), \tag{3.26}$$

$$(k_{\mathsf{min}}(n,\mu^*), m_{\mathsf{min}}(n,\mu^*)) \;:=\; \operatorname*{argmin}_i \, p^*_{n,k_i(n),m_i(n)}(\mu^*), \tag{3.27}$$

and find the value of $\lambda \in [0,1]$ for which

$$\lambda\Big(\sum_n p_{n,k_{\mathsf{min}}(n,\mu^*),m_{\mathsf{min}}(n,\mu^*)}(\mu^*)\Big) + (1-\lambda)\Big(\sum_n p_{n,k_{\mathsf{max}}(n,\mu^*),m_{\mathsf{max}}(n,\mu^*)}(\mu^*)\Big) = P_{\mathsf{con}}, \tag{3.28}$$

i.e.,

$$\lambda = \frac{\sum_n p_{n,k_{\mathsf{max}}(n,\mu^*),m_{\mathsf{max}}(n,\mu^*)}(\mu^*) - P_{\mathsf{con}}}{\sum_n p_{n,k_{\mathsf{max}}(n,\mu^*),m_{\mathsf{max}}(n,\mu^*)}(\mu^*) - \sum_n p_{n,k_{\mathsf{min}}(n,\mu^*),m_{\mathsf{min}}(n,\mu^*)}(\mu^*)}. \tag{3.29}$$

Now, defining two specific allocations, $\mathbf{I}^{\mathsf{min}}(\mu^*)$ and $\mathbf{I}^{\mathsf{max}}(\mu^*)$, as

$$I^{\mathsf{min}}_{n,k,m}(\mu^*) = \begin{cases} 1 & (k,m) = (k_{\mathsf{min}}(n,\mu^*), m_{\mathsf{min}}(n,\mu^*)) \\ 0 & \text{otherwise,} \end{cases} \text{ and}$$

$$I^{\mathsf{max}}_{n,k,m}(\mu^*) = \begin{cases} 1 & (k,m) = (k_{\mathsf{max}}(n,\mu^*), m_{\mathsf{max}}(n,\mu^*)) \\ 0 & \text{otherwise,} \end{cases} \tag{3.30}$$

respectively, the optimal user-MCS allocation is given by $\mathbf{I}^*_{\mathsf{CSRA}} = \lambda \mathbf{I}^{\mathsf{min}}(\mu^*) + (1-\lambda)\mathbf{I}^{\mathsf{max}}(\mu^*)$. The corresponding optimal power allocation is then given by (3.22). It can be seen that this solution satisfies the subchannel constraint as well as the sum power constraint with equality, i.e.,

$$\sum_{n,k,m} I^*_{n,k,m}(\mu^*) p^*_{n,k,m}(\mu^*) = \sum_{n,k,m} x^*_{n,k,m}(\mu^*, \mathbf{I}^*(\mu^*)) = P_{\mathsf{con}}.$$

Two interesting observations can be made from the above discussion. Firstly, for any choice of concave utility functions $U_{n,k,m}(\cdot)$, there exists an optimal scheduling and resource allocation strategy that allocates each subchannel to at most 2 user-MCS combinations. Therefore, when allocating N subchannels, even if more than $2N$ user-MCS options are available, at most $2N$ such options will be used. Secondly, if $\mathbf{I}^{\mathsf{min}}(\mu^*) = \mathbf{I}^{\mathsf{max}}(\mu^*)$, then the exact CSRA solution allocates power to at most one (k,m) combination for every subchannel, i.e., no subchannel is shared among any two or more user-MCS combinations. This observation will motivate the SRA problem's solution without subchannel sharing in Section 3.5.

3.4.4 Algorithmic implementation

In practice, it is not possible to search exhaustively over $\mu \in [\mu_{\mathsf{min}}, \mu_{\mathsf{max}}]$. Thus, we propose an algorithm to reach solutions in close (and adjustable) proximity to

the optimal. The algorithm first narrows down the location of μ^* using a bisection-search over $[\mu_{\mathsf{min}}, \mu_{\mathsf{max}}]$ for the optimum total power allocation, and then finds a set of resource allocations $(\mathbf{I}, \mathbf{x})$ that achieve a total utility close to the optimal.

To proceed in this direction, with the aim of developing a framework to do bisection-search over μ, let us define the total optimal allocated power for a given value of μ as follows:

$$X^*_{\mathsf{tot}}(\mu) \triangleq \sum_{n,k,m} x^*_{n,k,m}(\mu, \mathbf{I}^*(\mu)), \tag{3.31}$$

where $\mathbf{I}^*(\mu)$ and $\mathbf{x}^*(\mu, \mathbf{I}^*(\mu))$ (defined in (3.6)) minimize the Lagrangian (defined in (3.5)) for a given μ. The following lemma relates the variation of $X^*_{\mathsf{tot}}(\mu)$ with respect to μ.

Lemma 2. *The total optimal power allocation, $X^*_{\mathsf{tot}}(\mu)$, is a monotonically decreasing function of μ.*

Proof. Proof is given in Appendix B.2. □

A sample plot of $X^*_{\mathsf{tot}}(\mu)$ and $L(\mu, \mathbf{I}^*(\mu), \mathbf{x}^*(\mu, \mathbf{I}^*(\mu)))$ as a function of μ is shown in Figure 3.3. From the figure, three observations can be made. First, as μ increases, the optimal total allocated power decreases, as expected from Lemma 2. Second, as expected, the Lagrangian is maximized for that value of μ at which $X^*_{\mathsf{tot}}(\mu) = P_{\mathsf{con}}$. Third, the optimal total power allocation varies continuously in the region of μ where the optimal allocation, $\mathbf{I}^*(\mu)$, remains constant and takes a jump (negative) when $\mathbf{I}^*(\mu)$ changes. This happens for the following reason. We know, for any (n, k, m), that $p^*_{n,k,m}(\mu)$ is a continuous function of μ. Thus, when the optimal allocation remains constant over a range of μ, the total power allocated, $\sum_{n,k,m} I^*_{n,k,m}(\mu) p^*_{n,k,m}(\mu)$ also varies continuously with μ. However, at the point of discontinuity (say $\tilde{\mu}$), multiple optimal allocations achieve the same optimal value of Lagrangian. In other words,

$|S_n(\tilde{\mu})| > 1$ for some n. In that case, $X^*_{\mathsf{tot}}(\tilde{\mu})$ can take any value in the interval

$$\Big[\sum_n p^*_{n,k_{\mathsf{min}}(n),m_{\mathsf{min}}(n)}(\tilde{\mu}),\ \ \sum_n p^*_{n,k_{\mathsf{max}}(n),m_{\mathsf{max}}(n)}(\tilde{\mu})\Big]$$

while achieving the same minimum value of the Lagrangian at $\tilde{\mu}$. Applying Lemma 2, we have

$$X^*_{\mathsf{tot}}(\tilde{\mu}-\Delta_1) \ \geq\ \sum_n p^*_{n,k_{\mathsf{max}}(n),m_{\mathsf{max}}(n)}(\tilde{\mu}) \ \geq\ X^*_{\mathsf{tot}}(\tilde{\mu}) \ \geq\ \sum_n p^*_{n,k_{\mathsf{min}}(n),m_{\mathsf{min}}(n)}(\tilde{\mu}) \ \geq\ X^*_{\mathsf{tot}}(\tilde{\mu}+\Delta_2)$$

for any $\Delta_1, \Delta_2 > 0$, causing a jump of $\big(\sum_n p^*_{n,k_{\mathsf{min}}(n),m_{\mathsf{min}}(n)}(\tilde{\mu}) - \sum_n p^*_{n,k_{\mathsf{max}}(n),m_{\mathsf{max}}(n)}(\tilde{\mu})\big)$ in the total optimal power allocation at $\tilde{\mu}$.

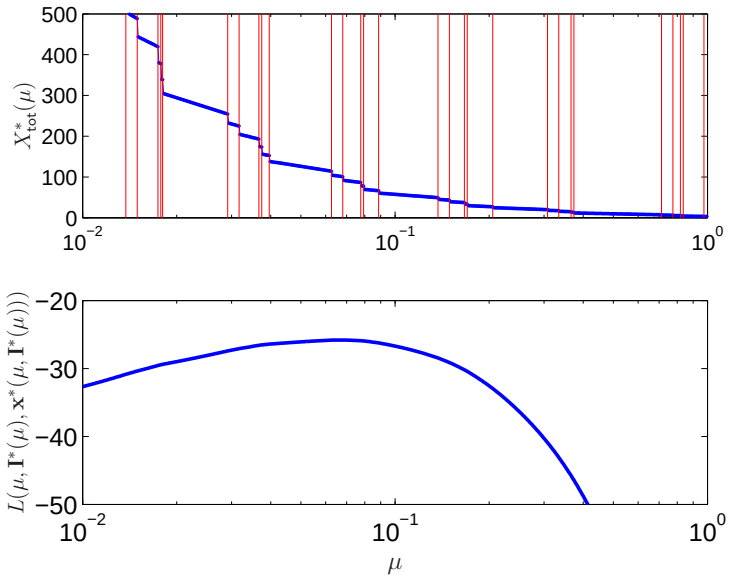

Figure 3.3: Prototypical plot of $X^*_{\mathsf{tot}}(\mu)$ and $L(\mu, \mathbf{I}^*(\mu), \mathbf{x}^*(\mu, \mathbf{I}^*(\mu)))$ as a function of μ for $N = K = 5$, and $P_{\mathsf{con}} = 100$. (See Section 3.6 for details.) The red vertical lines in the top plot show that a change in $\mathbf{I}^*(\mu)$ occurs at that μ.

Lemma 2 allows us to do a bisection-search over μ since $X^*_{\mathsf{tot}}(\mu)$ is a decreasing function of μ and the optimal μ is the one at which $X^*_{\mathsf{tot}}(\mu) = P_{\mathsf{con}}$. In particular,

if $\mu^* \in [\underline{\mu}, \bar{\mu}]$ for some $\underline{\mu}$ and $\bar{\mu}$, then $\mu^* \in \left[\frac{\underline{\mu}+\bar{\mu}}{2}, \bar{\mu}\right]$ if $X^*_{\mathsf{tot}}\left(\frac{\underline{\mu}+\bar{\mu}}{2}\right) > P_{\mathsf{con}}$, otherwise $\mu^* \in \left[\underline{\mu}, \frac{\underline{\mu}+\bar{\mu}}{2}\right]$. Using this concept, we propose an algorithm in Table 3.1 (see end of chapter) that finds an interval $[\underline{\mu}, \bar{\mu}]$, such that $\mu^* \in [\underline{\mu}, \bar{\mu}]$ and $\bar{\mu} - \underline{\mu} \leq \kappa$, where κ (> 0) is a tuning-parameter, and allocates resources based on optimal resource allocations at $\underline{\mu}$ and $\bar{\mu}$.

The following lemma characterizes the relationship between the tuning parameter κ and the accuracy of the obtained solution.

Lemma 3. *Let $\mu^* \in [\underline{\mu}, \bar{\mu}]$ be the point where the proposed CSRA algorithm stops, and the total utility obtained by the proposed algorithm and the exact CSRA solution be $\hat{U}_{\mathsf{CSRA}}(\underline{\mu}, \bar{\mu})$ and U^*_{CSRA}, respectively. Then, $0 \leq U^*_{\mathsf{CSRA}} - \hat{U}_{\mathsf{CSRA}}(\underline{\mu}, \bar{\mu}) \leq (\bar{\mu} - \underline{\mu})P_{\mathsf{con}}$.*

Proof. For proof, see Appendix B.3. □

Since our algorithm stops when $\bar{\mu} - \underline{\mu} \leq \kappa$, from Lemma 3, the gap between the obtained utility and the optimal utility is bounded by $P_{\mathsf{con}}\kappa$. Moreover, $\lim_{\underline{\mu} \to \bar{\mu}} \hat{U}_{\mathsf{CSRA}}(\underline{\mu}, \bar{\mu}) = U^*_{\mathsf{CSRA}}$.

The proposed algorithm requires at most $\left\lceil \log_2 \left(\frac{\mu_{\mathsf{max}} - \mu_{\mathsf{min}}}{\kappa} \right) \right\rceil$ iterations of μ in order to find $\bar{\mu}$, and $\underline{\mu}$ such that $\bar{\mu} - \underline{\mu} \leq \kappa$ and $\mu^* \in [\underline{\mu}, \bar{\mu}]$. Therefore, measuring the complexity of the algorithm by the number of times (3.11) must be solved for a given (n, k, m, μ), the proposed algorithm takes at most

$$NKM \left\lceil \log_2 \left(\frac{\mu_{\mathsf{max}} - \mu_{\mathsf{min}}}{\kappa} \right) \right\rceil \tag{3.32}$$

steps. We use this method of measuring complexity because it allows us to easily compare all algorithms in this chapter. Note that, for a given κ, the number of steps taken by the proposed bisection algorithm is proportional to $\log_2 \kappa$.

3.4.5 Some properties of the CSRA solution

In this subsection, we study a few properties of the CSRA solution that yield valuable insights into the optimal resource allocation strategy for any given value of Lagrange multiplier, μ. Let us fix a $\tilde{\mu} \in [\mu_{\mathsf{min}}, \mu_{\mathsf{max}}]$. Now, if $|S_n(\tilde{\mu})| \leq 1, \forall n$, then the optimal allocation at $\tilde{\mu}$, $\mathbf{I}^*(\tilde{\mu})$, is given by (3.16), which reveals that $\mathbf{I}^*(\tilde{\mu}) \in \{0,1\}^{N\times K\times M}$. In this case, the definition of $\mathcal{I}_{\mathsf{CSRA}}$ implies that every subchannel is allocated to at most one user-MCS combination. Note that this is precisely the constraint we impose in the later part of this chapter. Let us now consider the case where it is possible that $|S_n(\tilde{\mu})| > 1$ for some n.

Lemma 4. *For any $\tilde{\mu} > 0$, there exists a $\delta > 0$ such that for all $\mu \in (\tilde{\mu}-\delta, \tilde{\mu}+\delta)\backslash\{\tilde{\mu}\}$, there exists an optimal allocation, $\mathbf{I}^*(\mu) \in \mathcal{I}_{\mathsf{CSRA}}$, that satisfies $\mathbf{I}^*(\mu) \in \{0,1\}^{N\times K\times M}$. Moreover, if $\mu_1, \mu_2 \in (\tilde{\mu}-\delta, \tilde{\mu})$, then there exists $\mathbf{I}^*(\mu_1), \mathbf{I}^*(\mu_2) \in \{0,1\}^{N\times K\times M}$ such that $\mathbf{I}^*(\mu_1) = \mathbf{I}^*(\mu_2)$. The same property holds if both $\mu_1, \mu_2 \in (\tilde{\mu}, \tilde{\mu}+\delta)$.*

Proof. Proof is given in Appendix B.4. □

In conjunction with (3.12), the above lemma implies that the discontinuities in Fig. 3.3 are isolated and that, around every point on the horizontal axis, there is a small region over which $X^*_{\mathsf{tot}}(\mu)$ is continuous. Hence, the number of such discontinuities are, at most, countable.

3.5 Scheduling and Resource Allocation without subchannel sharing

In this section, we will solve the scheduling and resource allocation (SRA) problem (3.1) under the constraint that $I_{n,k,m} \in \{0,1\}$, i.e., that each subchannel can be allocated to at most one combination of user and MCS per time slot. We will refer to this problem as the "discrete scheduling and resource allocation" (DSRA) problem.

Storing the values of $I_{n,k,m}$ in the $N \times K \times M$ matrix $\mathbf{I}$, the DSRA subchannel constraint can be expressed as $\mathbf{I} \in \mathcal{I}_{\mathsf{DSRA}}$, where

$$\mathcal{I}_{\mathsf{DSRA}} := \left\{ \mathbf{I} : \mathbf{I} \in \{0,1\}^{N \times K \times M}, \sum_{k,m} I_{n,k,m} \leq 1 \ \forall n \right\}.$$

Then, using (3.1), the DSRA problem can be stated as

$$\begin{aligned} \text{DSRA} \ := \ & \max_{\substack{\{p_{n,k,m} \geq 0\} \\ \mathbf{I} \in \mathcal{I}_{\mathsf{DSRA}}}} \sum_{n,k,m} I_{n,k,m} \, \mathrm{E} \left\{ U_{n,k,m} \big((1 - a_{k,m} e^{-b_{k,m} p_{n,k,m} \gamma_{n,k}}) r_{k,m} \big) \right\} \\ & \text{s.t.} \sum_{n,k,m} I_{n,k,m} p_{n,k,m} \leq P_{\mathsf{con}}. \end{aligned} \tag{3.33}$$

Let us denote the optimal $\mathbf{I}$ and $\mathbf{p}$ for (3.33) by $\mathbf{I}^*_{\mathsf{DSRA}}$ and $\mathbf{p}^*_{\mathsf{DSRA}}$, respectively.

The DSRA problem is a mixed-integer programming problem. Mixed-integer programming problems are generally NP-hard, meaning that polynomial-time solutions do not exist [43]. Fortunately, in some cases, such as ours, one can exploit the problem structure to design polynomial-complexity algorithms that reach solutions in close vicinity of the exact solution. We first describe an approach to solve the DSRA mixed-integer programming problem exactly by exhaustively searching over all possible user-MCS allocations in order to arrive at the optimal user, rate, and power allocation. We will see that this "brute-force" approach has a complexity that grows exponentially in the number of subchannels. Later, we will exploit the DSRA problem structure, and its relation to the CSRA problem, to design an algorithm with near-optimal performance and polynomial complexity.

3.5.1 Brute-force algorithm

Consider that, if we attempted to solve our DSRA problem via brute-force (i.e., by solving the power allocation sub-problem for every possible choice of $\mathbf{I} \in \mathcal{I}_{\mathsf{DSRA}}$),

we would solve the following sub-problem for every given $\mathbf{I}$.

$$\max_{\{p_{n,k,m}\geq 0\}} \sum_{n,k,m} I_{n,k,m}\, \mathrm{E}\left\{U_{n,k,m}\big((1-a_{k,m}e^{-b_{k,m}p_{n,k,m}\gamma_{n,k}})r_{k,m}\big)\right\}$$
$$\text{s.t.} \sum_{n,k,m} I_{n,k,m}\, p_{n,k,m} \leq P_{\mathsf{con}}. \tag{3.34}$$

Borrowing our approach to the CSRA problem, we could transform the variable $p_{n,k,m}$ into $x_{n,k,m}$ via the relation: $x_{n,k,m} = I_{n,k,m}\, p_{n,k,m}$. The problem in (3.34) can, therefore, be written as:

$$\min_{\{x_{n,k,m}\geq 0\}} \sum_{n,k,m} I_{n,k,m}\, F_{n,k,m}(I_{n,k,m}, x_{n,k,m}) \;\; \text{s.t.} \sum_{n,k,m} x_{n,k,m} \leq P_{\mathsf{con}}, \tag{3.35}$$

where $F_{n,k,m}(I_{n,k,m}, x_{n,k,m})$ is defined in (3.4). This problem is a convex optimization problem that satisfies Slater's condition [42] when $x_{n,k,m} = P_{\mathsf{con}}/2NKM$ for all n, k, m. Therefore, its solution is equal to the solution of its dual problem (i.e., zero duality gap) [42]. To formulate the dual problem, we write the Lagrangian of the primal problem (3.35) as

$$L_{\mathbf{I}}(\mu, \mathbf{x}) = \sum_{n,k,m} I_{n,k,m} F_{n,k,m}(I_{n,k,m}, x_{n,k,m}) + \Big(\sum_{n,k,m} x_{n,k,m} - P_{\mathsf{con}}\Big)\mu, \tag{3.36}$$

where μ is the dual variable and $\mathbf{x}$ is the $N \times K \times M$ matrix containing actual powers allocated to all (n, k, m) combinations. Note that the Lagrangian in (3.36) is exactly the same as the Lagrangian for the CSRA problem in (3.5). Using (3.36), the dual of the brute-force problem can be written as

$$\max_{\mu\geq 0} \min_{\mathbf{x}\succeq 0} L_{\mathbf{I}}(\mu, \mathbf{x}) = \max_{\mu\geq 0} L_{\mathbf{I}}(\mu, \mathbf{x}^*(\mu)) = L_{\mathbf{I}}(\mu^*_{\mathbf{I}}, \mathbf{x}^*(\mu^*_{\mathbf{I}})), \tag{3.37}$$

for optimal solutions $\mu^*_{\mathbf{I}}$ and $\mathbf{x}^*(\mu^*_{\mathbf{I}})$. Minimizing $L_{\mathbf{I}}(\mu, \mathbf{x})$ over $\{\mathbf{x} \succeq 0\}$ by equating the differential of $L_{\mathbf{I}}(\mu, \mathbf{x})$ w.r.t. $x_{n,k,m}$ to zero (which is identical to the approach taken in Section 3.4.1 for the CSRA problem), we get that, for any subchannel n,

$$x^*_{n,k,m}(\mu) = I_{n,k,m}\, p^*_{n,k,m}(\mu). \tag{3.38}$$

Here,

$$p^*_{n,k,m}(\mu) = \begin{cases} \tilde{p}_{n,k,m}(\mu) & \text{if } 0 \le \mu \le a_{k,m} b_{k,m} r_{k,m} U'_{n,k,m}\big((1-a_{k,m}) r_{k,m}\big)\, \mathrm{E}\{\gamma_{n,k}\} \\ 0 & \text{otherwise,} \end{cases} \tag{3.39}$$

and $\tilde{p}_{n,k,m}(\mu)$ is the unique[13] value satisfying (3.11), repeated as (3.40) for convenience.

$$\mu = a_{k,m} b_{k,m} r_{k,m}\, \mathrm{E}\left\{ U'_{n,k,m}\big((1 - a_{k,m} e^{-b_{k,m}\tilde{p}_{n,k,m}(\mu)\gamma_{n,k}}) r_{k,m}\big) \gamma_{n,k} e^{-b_{k,m}\tilde{p}_{n,k,m}(\mu)\gamma_{n,k}} \right\}. \tag{3.40}$$

Note that the Lagrangian as well as the power allocation in (3.36) and (3.38) are identical to that obtained for the CSRA problem in (3.5) and (3.12), respectively. Also recall that (3.19)-(3.21) hold even when $\mathbf{I}^*(\mu)$ is replaced by arbitrary $\mathbf{I}$. Thus, we have $\mu^*_{\mathbf{I}} \in [\mu_{\mathsf{min}}, \mu_{\mathsf{max}}]$, where μ_{min} and μ_{min} are defined in (3.20) and (3.21), respectively.

As discussed in Section 3.4.1, $\tilde{p}_{n,k,m}(\mu)$ is a strictly-decreasing continuous function of μ, which makes $p^*_{n,k,m}(\mu)$ a decreasing continuous function of μ. Let us now define

$$X^*_{\mathsf{tot}}(\mathbf{I}, \mu) \triangleq \sum_{n,k,m} x^*_{n,k,m}(\mu) = \sum_{n,k,m} I_{n,k,m} p^*_{n,k,m}(\mu) \tag{3.41}$$

as the total optimal power allocation for allocation $\mathbf{I}$ at μ. Therefore, $X^*_{\mathsf{tot}}(\mathbf{I}, \mu)$ is also a decreasing continuous function of μ. This reduces our problem to finding the minimum value of $\mu \in [\mu_{\mathsf{min}}, \mu_{\mathsf{max}}]$ for which $X^*_{\mathsf{tot}}(\mathbf{I}, \mu) = P_{\mathsf{con}}$. Such a problem structure (i.e., finding the minimum Lagrange multiplier satisfying a sum-power constraint) yields a *water-filling* solution (e.g., [33, 44]). To obtain such a solution (in our case, $\mu^*_{\mathbf{I}}$) one can use the bisection-search algorithm given in Table 3.1 (see end of chapter).

While there are many ways to find μ, we focus on bisection-search for easy comparison to the CSRA algorithm. Then, to solve the resource allocation problem for a

[13]By assumption, $U'_{n,k,m}(\cdot)$ is a decreasing positive function and $e^{-b_{k,m}\tilde{p}_{n,k,m}(\mu)\gamma_{n,k}}$ is a strictly-decreasing positive function of $\tilde{p}_{n,k,m}(\mu)$, which makes the right side of (3.40) a strictly-decreasing positive function of $\tilde{p}_{n,k,m}(\mu)$.

given $\mathbf{I} \in \mathcal{I}_{\mathsf{DSRA}}$, the complexity, in terms of the number of times (3.40) (or (3.11)) is solved to yield $\hat{\mu}_{\mathbf{I}}$ such that $|\hat{\mu}_{\mathbf{I}} - \mu^*_{\mathbf{I}}| < \kappa$, is $\left(\sum_{n,k,m} I_{n,k,m}\right) \left\lceil \log_2 \left(\frac{\mu_{\mathsf{max}} - \mu_{\mathsf{min}}}{\kappa}\right) \right\rceil$. Since the brute-force algorithm examines $|\mathcal{I}_{\mathsf{DSRA}}| = (KM+1)^N$ hypotheses of $\mathbf{I}$, the corresponding complexity needed to find the exact DSRA solution is $\left\lceil \log_2 \left(\frac{\mu_{\mathsf{max}} - \mu_{\mathsf{min}}}{\kappa}\right) \right\rceil \times \sum_{n=1}^{N} n\binom{N}{n}(KM)^n$ or, equivalently,

$$\left\lceil \log_2 \left(\frac{\mu_{\mathsf{max}} - \mu_{\mathsf{min}}}{\kappa}\right) \right\rceil \times (KM+1)^{N-1} NKM. \tag{3.42}$$

Because this "brute-force" algorithm may be impractical to implement for practical values of K, M, and N, we focus, in the sequel, on lower-complexity DSRA approximations. In doing so, we exploit insights previously gained from our study of the CSRA problem.

3.5.2 Proposed DSRA algorithm

Equation (3.30) in Section 3.4.2 demonstrated that there exists an optimal user-MCS allocation for the CSRA problem that either lies in the domain of DSRA problem, i.e., $\mathbf{I}^*(\mu^*) \in \mathcal{I}_{\mathsf{DSRA}}$, or is a convex combination of two points from the domain of DSRA problem, i.e., $\mathbf{I}^*(\mu^*) = \lambda \mathbf{I}^{\mathsf{min}}(\mu^*) + (1-\lambda)\mathbf{I}^{\mathsf{max}}(\mu^*)$, where $\mathbf{I}^{\mathsf{min}}(\mu^*) \neq \mathbf{I}^{\mathsf{max}}(\mu^*)$ and $\mathbf{I}^{\mathsf{min}}(\mu^*), \mathbf{I}^{\mathsf{max}}(\mu^*) \in \mathcal{I}_{\mathsf{DSRA}}$. (Note that if $\mathbf{I} \in \mathcal{I}_{\mathsf{CSRA}}$ and $\mathbf{I} \in \{0,1\}^{N \times K \times M}$, then $\mathbf{I} \in \mathcal{I}_{\mathsf{DSRA}}$.) This observation motivates us to attack the DSRA problem using the CSRA algorithm. In this section, we provide the details of such an approach.

The following lemma will be instrumental in understanding the relationship between the CSRA and DSRA problems and will serve as the basis for allocating resources in the DSRA problem setup.

Lemma 5. *If the solution of the Lagrangian dual of the CSRA problem (3.6) for a given μ is such that $\mathbf{I}^*(\mu) \in \{0,1\}^{N \times K \times M}$, and the corresponding total power is*

$X^*_{\mathsf{tot}}(\mu)$ *as in (3.31), then the solution to the optimization problem*

$$\begin{aligned}(\mathbb{P}^*, \mathbb{I}^*) \;=\; & \underset{\substack{\{\mathbb{P}\succeq 0\}\\ \mathbf{I}\in\mathcal{I}_{\mathsf{DSRA}}}}{\operatorname{argmax}} \sum_{n,k,m} \mathbb{I}_{n,k,m}\, \mathrm{E}\left\{U_{n,k,m}\big((1-a_{k,m}e^{-b_{k,m}\mathbb{P}_{n,k,m}\gamma_{n,k}})r_{k,m}\big)\right\} \\ & \text{s.t.} \sum_{n,k,m} \mathbb{I}_{n,k,m}\mathbb{P}_{n,k,m} \le X^*_{\mathsf{tot}}(\mu)\end{aligned}$$

satisfies $\mathbb{I}^* = \mathbf{I}^*(\mu)$ *and, for every* (n,k,m), $\mathbb{P}^*_{n,k,m} = \begin{cases} \frac{x^*_{n,k,m}(\mu,\mathbf{I}^*(\mu))}{I^*_{n,k,m}(\mu)} & \text{if } I^*_{n,k,m}(\mu)\neq 0 \\ 0 & \text{otherwise.}\end{cases}$

Proof. Proof is given in Appendix B.5. □

From the above lemma, we conclude that if a μ exists such that $\mathbf{I}^*(\mu) \in \mathcal{I}_{\mathsf{DSRA}}$ and $X^*_{\mathsf{tot}}(\mu) = P_{\mathsf{con}}$, then the DSRA problem is solved exactly by the CSRA solution $(\mathbf{I}^*(\mu), \mathbf{x}^*(\mu, \mathbf{I}^*(\mu)))$, i.e., the optimal user-MCS allocation $\mathbf{I}^*_{\mathsf{DSRA}}$ equals $\mathbf{I}^*(\mu)$ and the optimal power allocation, $\mathbf{p}^*_{\mathsf{DSRA}}$, for any (n,k,m), is

$$p^*_{n,k,m,\,\mathsf{DSRA}} = \begin{cases} \frac{x^*_{n,k,m}(\mu,\mathbf{I}^*(\mu)))}{I^*_{n,k,m}(\mu)} & \text{if } I^*_{n,k,m}(\mu)\neq 0 \\ 0 & \text{otherwise.}\end{cases} \tag{3.43}$$

Recall that the optimal total power achieved for a given value of Lagrange multiplier μ, i.e., $X^*_{\mathsf{tot}}(\mu) = \sum_{n,k,m} x^*_{n,k,m}(\mu, \mathbf{I}^*(\mu))$, is piece-wise continuous and that a discontinuity (or "gap") occurs at μ when multiple allocations achieving the same optimal value of Lagrangian exist. When the sum-power constraint, P_{con}, lies in one of those "gaps," the optimal allocation for the CSRA problem equals a convex combination of two elements from the set $\mathcal{I}_{\mathsf{DSRA}}$, and the CSRA solution is not admissible for DSRA. In such cases, we are motivated to choose the approximate DSRA solution $\hat{\mathbf{I}}_{\mathsf{DSRA}} \in \{\mathbf{I}^{\mathsf{min}}(\mu), \mathbf{I}^{\mathsf{max}}(\mu)\}$ yielding highest utility. In Table 3.1 (see end of chapter), we detail an implementation of our proposed DSRA algorithm that has significantly lower complexity than brute-force. The numerical simulations in Section 3.6 show that its performance is very close to optimal. Moreover, the following lemma bounds

the asymptotic difference in utility achieved by the exact DSRA solution and that produced by our proposed DSRA algorithm.

Lemma 6. *Let μ^* be the optimal μ for the CSRA problem and $\underline{\mu}, \bar{\mu}$ be such that $\mu^* \in [\underline{\mu}, \bar{\mu}]$. Let U^*_{DSRA} and $\hat{U}_{\mathsf{DSRA}}(\underline{\mu}, \bar{\mu})$ be the utilities achieved by the exact DSRA solution and the proposed DSRA algorithm, respectively. Then,*

$$\begin{aligned} 0 \;&\leq\; U^*_{\mathsf{DSRA}} - \lim_{\underline{\mu} \to \bar{\mu}} \hat{U}_{\mathsf{DSRA}}(\underline{\mu}, \bar{\mu}) \\ &\leq\; (\mu^* - \mu_{\mathsf{min}})\big(P_{\mathsf{con}} - X^*_{\mathsf{tot}}(\mathbf{I}^{\mathsf{min}}(\mu^*), \mu^*)\big) \qquad (3.44) \\ &\leq\; \begin{cases} 0 & \text{if } |S_n(\mu^*)| \leq 1 \;\forall n \\ \big(\mu_{\mathsf{max}} - \mu_{\mathsf{min}}\big) P_{\mathsf{con}} & \text{otherwise} \end{cases}. \qquad (3.45) \end{aligned}$$

Proof. The proof is given in Appendix B.6. □

It is interesting to note that the bound (3.45) does not scale with number of users K or subchannels N.

The complexity of the proposed DSRA algorithm is marginally greater than that of the CSRA algorithm, since an additional comparison of two possible user-MCS allocation choices is involved. In units of solving (3.11) for a given (n, k, m, μ), the DSRA complexity is at most

$$N(KM + 2) \left\lceil \log_2 \left(\tfrac{\mu_{\mathsf{max}} - \mu_{\mathsf{min}}}{\kappa} \right) \right\rceil. \qquad (3.46)$$

Comparing (3.42) and (3.46), we find that the complexity of the proposed DSRA algorithm is polynomial in N, K, M, which is considerably less than that of the brute-force algorithm (i.e., exponential in N).

3.5.3 Discussion

Before concluding this section, we make some remarks about our approach to DSRA and its connections to CSRA. First, we note that the DSRA problem is an integer-programming problem due to the discrete domain $\{0, 1\}$ assumed for $I_{n,k,m}$.

Because integer programming problems are generally NP-hard (recall our "brute force" DSRA solution), one is strongly motivated to find a polynomial-complexity method whose performance is as high as possible. One possible approach is based on "relaxation," whereby the discrete domain is relaxed to an interval domain, the relaxed problem is solved (with polynomial complexity), and the obtained solution is mapped back to the discrete domain. In fact, relaxation was previously employed for OFDMA frequency-scheduling in [33, 35], and the DSRA approximation that we propose in Section 3.5.2 can also be interpreted as a form of relaxation.

The optimization literature suggests that relaxation is successful in some—but not all—cases, implying that relaxation-based OFDMA algorithms must be designed with care. For example, relaxation has widely used to solve *linear* integer programs (LIPs) [45–47]. The DSRA, however, is a mixed-integer nonlinear program (MINLP), and for such problems relaxation does not always perform well [47,48]. Now, one could cite the analysis in [49, p. 371], which shows that—for a broad class of integer programming problems—the duality gap goes to zero as the number of integer variables goes to infinity, to suggest that the DSRA problem can be well approximated by its relaxed counterpart, CSRA, as the number of OFDMA subchannels $N \to \infty$. However, in practice, the number of subchannels N is often quite small, preventing the application of this argument. For example, in LTE systems [30, 50], each subchannel consists of 12 subcarriers, so that only 25 subchannels are used for 5 MHz bandwidths, and only 6 are used for 1.4 MHz bandwidths.

The above considerations have motivated us to investigate, in detail, the relationship between the continuous and discrete resource allocation scenarios. The results of

our investigation include insights into the dissimilarity between CSRA and DSRA solutions (e.g., Lemma 4 and Lemma 5), and an efficient polynomial-complexity DSRA approximation that (as we shall see in Section 3.6) performs near-optimally for all N and admits the tight performance bound (3.45).

3.6 Numerical Evaluation

In this section, we analyze the performance of an OFDMA downlink system that uses the proposed CSRA and DSRA algorithms for scheduling and resource allocation under different system parameters. Unless otherwise specified, we use the sum-goodput utility $U_{n,k,m}(g) = g$.

For downlink transmission, the BS employs an uncoded 2^{m+1}-QAM signaling scheme with MCS index $m \in \{1, \ldots, 15\}$. In this case, we have $r_{k,m} = m + 1$ bits per symbol and one symbol per codeword. In the error rate model $\epsilon_{k,m}(p\gamma) = a_{k,m}e^{-b_{k,m}p\gamma}$, we choose $a_{k,m} = 1$ and $b_{k,m} = 1.5/(2^{m+1} - 1)$ because the actual symbol error rate of a 2^{m+1}-QAM system is proportional to $\exp(-1.5p\gamma/(2^{m+1} - 1))$ in the high-$(p\gamma)$ regime [51] and is ≈ 1 when $p\gamma = 0$. We use the standard OFDM model [52] to describe the (instantaneous) frequency-domain observation made by the k^{th} user on the n^{th} subchannel:

$$y_{n,k} = h_{n,k}x_n + \nu_{n,k}, \quad \text{for } n \in \{1, \ldots, N\} \text{ and } k \in \{1, \ldots, K\} \tag{3.47}$$

In (3.47), x_n denotes the QAM symbol broadcast by the BS on the n^{th} subchannel, $h_{n,k}$ the gain of the n^{th} subchannel between the k^{th} user and the BS, and $\nu_{n,k}$ a corresponding complex Gaussian noise sample. We assume that $\{\nu_{n,k}\}$ is unit variance and white across (n, k), and we recall that the exogenous subchannel-SNR satisfies $\gamma_{n,k} = |h_{n,k}|^2$. We furthermore assume that the k^{th} user's frequency-domain channel

gains $\mathbf{h}_k = (h_{1,k}, \ldots, h_{N,k})^T \in \mathbb{C}^N$ are related to its channel impulse response $\mathbf{g}_k = (g_{1,k}, \ldots, g_{L,k})^T \in \mathbb{C}^L$ via $\mathbf{h}_k = \mathbf{F}\mathbf{g}_k$, where $\mathbf{F} \in \mathbb{C}^{N \times L}$ contains the first $L(< N)$ columns of the N-DFT matrix, and where $\{g_{l,k}\}$ are i.i.d. over (l, k) and drawn from a zero-mean complex Gaussian distribution with variance σ_g^2 chosen so that $\mathrm{E}\{\gamma_{n,k}\} = 1$. Since the total available power for all subchannels at the base-station is P_{con}, the average available SNR per subchannel will be denoted by $\mathsf{SNR} = \frac{P_{\mathsf{con}}}{N}\,\mathrm{E}\{\gamma_{n,k}\}$.

To model imperfect CSI, we assume that there is a channel-estimation period during which the mobiles take turns to each broadcast one pilot OFDM symbol, from which the BS estimates the corresponding subchannel gains. Furthermore, we assume that the channels do not vary between pilot and data periods. To estimate $\mathbf{h}_k$, we assume that the BS observes $\tilde{\boldsymbol{y}}_k = \sqrt{p_{\mathsf{pilot}}}\,\mathbf{h}_k + \tilde{\boldsymbol{\nu}}_k \in \mathbb{C}^N$. Note that the average SNR per subchannel under pilot transmission is $\mathsf{SNR}_{\mathsf{pilot}} = p_{\mathsf{pilot}}\,\mathrm{E}\{\gamma_{n,k}\}$. The channel $\mathbf{h}_k$ and the pilot observations $\tilde{\boldsymbol{y}}_k$ are zero-mean jointly Gaussian, and furthermore $\mathbf{h}_k|\tilde{\boldsymbol{y}}_k$ is Gaussian with mean $\mathrm{E}\{\mathbf{h}_k|\tilde{\boldsymbol{y}}_k\} = \mathbf{R}_{\mathbf{h}_k,\tilde{\boldsymbol{y}}_k}\mathbf{R}^{-1}_{\tilde{\boldsymbol{y}}_k,\tilde{\boldsymbol{y}}_k}\tilde{\boldsymbol{y}}_k$ and covariance $\mathrm{Cov}(\mathbf{h}_k|\tilde{\boldsymbol{y}}_k) = \mathbf{R}_{\mathbf{h}_k,\mathbf{h}_k} - \mathbf{R}_{\mathbf{h}_k,\tilde{\boldsymbol{y}}_k}\mathbf{R}^{-1}_{\tilde{\boldsymbol{y}}_k,\tilde{\boldsymbol{y}}_k}\mathbf{R}_{\tilde{\boldsymbol{y}}_k\mathbf{h}_k}$, where $\mathbf{R}_{\mathbf{z}_1,\mathbf{z}_2}$ denotes the cross-correlation of random vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ [53, pp. 155]. Since $\mathbf{R}_{\mathbf{h}_k,\mathbf{h}_k} = \sigma_g^2\mathbf{F}\mathbf{F}'$, $\mathbf{R}_{\mathbf{h}_k,\tilde{\boldsymbol{y}}_k} = \sqrt{p_{\mathsf{pilot}}}\sigma_g^2\mathbf{F}\mathbf{F}'$, and $\mathbf{R}_{\tilde{\boldsymbol{y}}_k,\tilde{\boldsymbol{y}}_k} = p_{\mathsf{pilot}}\sigma_g^2\mathbf{F}\mathbf{F}' + \mathbf{I}$ (where $\mathbf{I}$ denotes the identity matrix), it is straightforward to show that the elements on the diagonal of $\mathrm{Cov}(\mathbf{h}_k|\tilde{\boldsymbol{y}}_k)$ are equal. Furthermore, $\mathrm{E}\{\mathbf{h}_k|\tilde{\boldsymbol{y}}_k\}$ can be recognized as the pilot-aided MMSE estimate of $\mathbf{h}_k$. In summary, conditioned on the pilot observations, $h_{n,k}$ is Gaussian with mean $\hat{h}_{n,k}$ given by the n^{th} element of $\mathrm{E}\{\mathbf{h}_k|\tilde{\boldsymbol{y}}_k\}$, and with variance σ_e^2 given by the first diagonal element of $\mathrm{Cov}(\mathbf{h}_k|\tilde{\boldsymbol{y}}_k)$. Thus, conditioned on the pilot observations, $\gamma_{n,k}$ has a non-central chi-squared distribution with two degrees of freedom.

We will refer to the proposed CSRA and DSRA algorithms implemented under imperfect CSI as "CSRA-ICSI" and "DSRA-ICSI," respectively. Their performances will be compared to that of "CSRA-PCSI," i.e., CSRA implemented under perfect CSI, which serves as a performance upper bound, and *fixed-power random-user scheduling* (FP-RUS), which serves as a performance lower bound. FP-RUS schedules, on each subchannel, one user selected uniformly from $\{1, \ldots, K\}$, to which it allocates power P_{con}/N and the fixed MCS m that maximizes expected goodput. Unless specified, the number of OFDM subchannels is $N = 64$, the number of users is $K = 16$, the impulse response length is $L = 2$, the average SNR per subchannel is $\mathsf{SNR} = 10$ dB, the pilot SNR is $\mathsf{SNR}_{\mathsf{pilot}} = -10$ dB, and the DSRA/CSRA tuning parameter is $\kappa = 0.3/P_{\mathsf{con}}$ (recall Table 3.1 at the end of this chapter). In all plots, goodput values were empirically averaged over 1000 realizations.

Figure 3.4 plots the subchannel-averaged goodput achieved by the above-described scheduling and resource-allocation schemes for different grades of CSI. In this curve, $\mathsf{SNR}_{\mathsf{pilot}}$ is varied so as to obtain estimates of subchannel SNR with different grades of accuracy. All other parameters remain unchanged. The plot shows that, as $\mathsf{SNR}_{\mathsf{pilot}}$ is increased, the performance of the proposed schemes (under the availability of imperfect CSI) increases from that of FP-RUS to that achieved by CSRA-PCSI. This is expected because, with increasing $\mathsf{SNR}_{\mathsf{pilot}}$, the BS uses more accurate channel-state information for scheduling and resource allocation, and thus achieves higher goodput. The plot also shows that, even though the proposed CSRA algorithm exactly solves the CSRA problem and the proposed DSRA algorithm approximately solves the DSRA problem, their performances almost coincide. In particular, although the goodput achieved by CSRA-ICSI scheme exceeded that of DSRA-ICSI scheme in

up-to 49% of the realizations, the maximum difference in the subchannel-averaged goodput was merely 4×10^{-3} bits per channel-use (bpcu). Since the DSRA-ICSI schemes cannot achieve a sum-goodput higher than that achieved by the CSRA-ICSI scheme, it can be deduced that the proposed DSRA algorithm is exhibiting near-optimal performance.

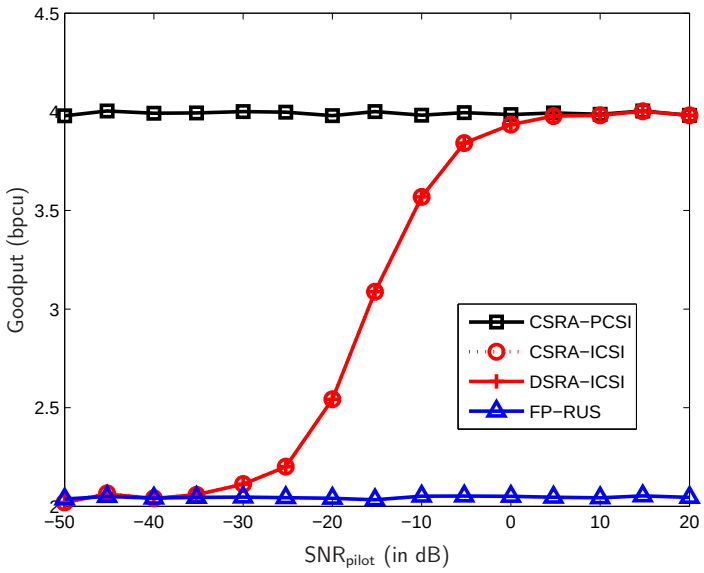

Figure 3.4: Average goodput per subchannel versus $\mathsf{SNR}_{\mathsf{pilot}}$. Here, $N = 64$, $K = 16$, and $\mathsf{SNR} = 10$ dB.

Figure 3.5 plots the subchannel-averaged goodput versus the number of available users, K, ranging between 1 and 32. It shows that, as K increases, the goodput per subchannel achieved by the proposed schemes increase under both perfect and imperfect CSI, whereas that achieved by the FP-RUS scheme remains constant. This is because, in the former case, the availability of more users can be exploited to schedule users with stronger subchannels, whereas, in the FP-RUS scheme, users are

scheduled without regard to the instantaneous channel conditions. Similar to the observations in the previous plots, the performance difference between the proposed CSRA and DSRA algorithms remains negligible. In particular, although the goodput achieved by CSRA-ICSI exceeded that of DSRA-ICSI in up-to 29% of the realizations, the maximum difference in the subchannel-averaged goodput was merely 7×10^{-4} bpcu.

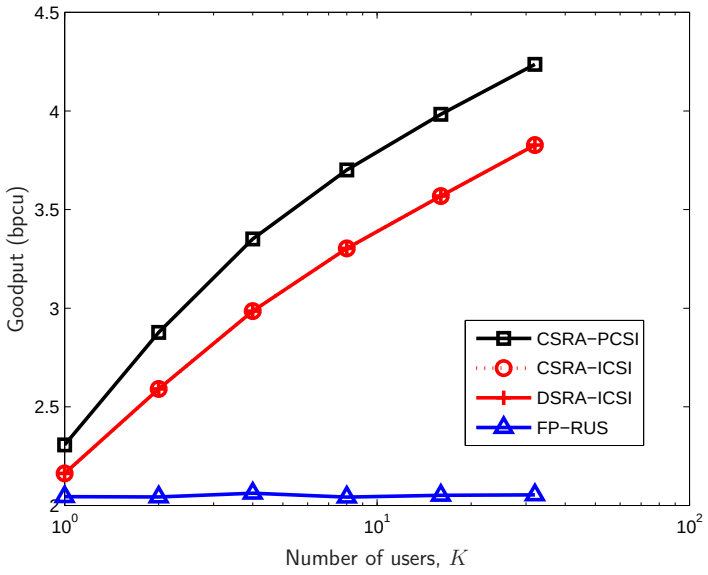

Figure 3.5: Average goodput per subchannel versus number of users, K. In this plot, $N = 64$, $\mathsf{SNR} = 10$ dB, and $\mathsf{SNR}_{\mathsf{pilot}} = -10$ dB.

In Figure 3.6, the top plot shows the subchannel-averaged goodput and the bottom plot shows the subchannel and realization-averaged value of the bound (3.44) on the DSRA-ICSI optimality gap as a function of SNR. In the top plot, it can be seen that, as SNR increases, the difference between CSRA-PCSI and CSRA-ICSI (or, DSRA-ICSI) increases. However, the difference grows slower than the difference between

CSRA-PCSI and FP-RUS. Interestingly, even for high values of SNR, the performance of CSRA-ICSI and DSRA-ICSI remain almost identical. In particular, although the goodput achieved by CSRA-ICSI scheme exceeded that of DSRA-ICSI scheme in up-to 28% of the realizations, the maximum difference in the subchannel-averaged goodput was merely 4×10^{-5} bpcu. The bottom plot, which illustrates the average value of $(\mu^* - \mu_{\mathsf{min}})\big(P_{\mathsf{con}} - X^*_{\mathsf{tot}}(\mathbf{I}^{\mathsf{min}}, \mu^*)\big)$ over all realizations and subchannels w.r.t. SNR, shows that the loss in sum-goodput over all subchannels due to the sub-optimality of proposed DSRA solution under imperfect CSI is bounded by 7×10^{-3} bpcu, even when the subchannel-averaged goodput of DSRA-ICSI is of the order of tens of bpcu. These results confirm that the bound (3.44) is quite tight at high SNR.

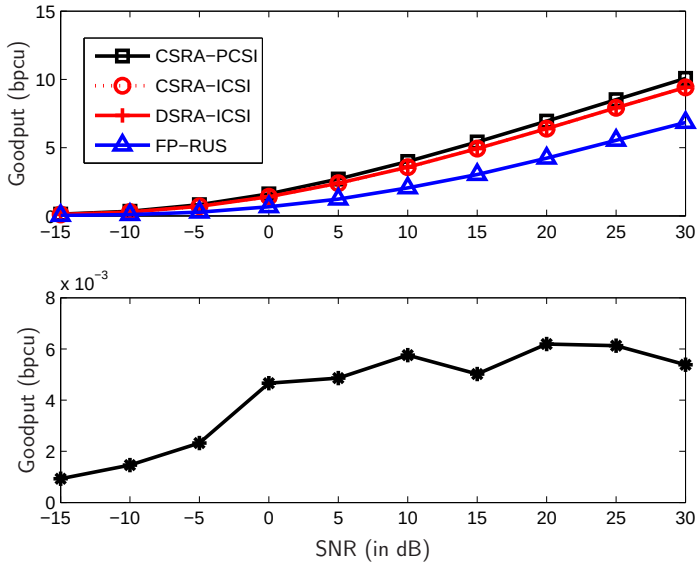

Figure 3.6: The top plot shows the average goodput per subchannel as a function of SNR. The bottom plot shows the average bound on the optimality gap between the proposed and exact DSRA solutions (given in (3.44)), i.e., the average value of $(\mu^* - \mu_{\min})(P_{\mathsf{con}} - X^*_{\mathsf{tot}}(\mathbf{I}^{\min}, \mu^*))/N$. In this plot, $N = 64$, $K = 16$, and $\mathsf{SNR}_{\mathsf{pilot}} = -10$ dB.

Figure 3.7 shows the performance of the proposed DSRA algorithm under a sum-utility criterion that is motivated by a common pricing model for an elastic application such as file-transfer [27, 28]. In particular, we partitioned the $K = 16$ users into two classes: $k \in \{1, \ldots, 8\} \triangleq \mathcal{K}_1$ is "Class 1" and $k \in \{9, \ldots, 16\} \triangleq \mathcal{K}_2$ is "Class 2," and we ran DSRA with the utility $U_k(g) \triangleq (1 - e^{-w_1 g})\mathbf{1}_{k \in \mathcal{K}_1} + (1 - e^{-w_2 g})\mathbf{1}_{k \in \mathcal{K}_2}$, where $\mathbf{1}_{\mathcal{E}}$ denotes the indicator of event $\mathcal{E}$. The utility can be regarded as the revenue earned by the operator: when $w_i > w_j$, Class-i users pay more (for a given goodput g) than Class-j users in exchange for priority service. In Fig. 3.7, we show the resulting DSRA-maximized utility summed over all users, as well as that summed over each individual user class. For comparison, we show the utility (summed over all users) when DSRA is "naively" used to maximize sum-*goodput* instead of sum-utility. The top plot in Fig. 3.7 shows performance as a function of w_1, for fixed $w_2 = 1$ and $\mathsf{SNR} = 0$ dB. There the behavior is as expected: when $w_1 \ll w_2 = 1$ (i.e., Class-1 users pay much less) DSRA allocates the overwhelming majority of the resources to Class-2 users, in an effort to earn more revenue. Meanwhile, when $w_1 \gg w_2 = 1$, the overwhelming majority of resources are allocated to Class-1 users. Moreover, it is evident that the naive goodput-maximizing scheme does not earn the operator as much revenue as the utility-maximizing scheme (outside of the trivial case that $w_1 = w_2$). The bottom plot in Fig. 3.7 shows the above described sum-utilities as a function of SNR, for fixed $w_1 = 0.85$ and $w_2 = 1$. There it can be seen that, at low SNR, the two classes achieve proportional utilities while, at high SNR, the utility of Class-1 users tend to zero. This behavior can be explained as follows: At low SNR, the goodputs g are small, in which case $1 - e^{-w_i g} \approx w_i g$, so that $U_k(g) \approx w_i g \mathbf{1}_{k \in \mathcal{K}_i}$, i.e., weighted-goodput utility. At high SNR, this approximation does not hold because

the goodputs g are usually large, and this particular pricing-based utility becomes increasingly unfair.

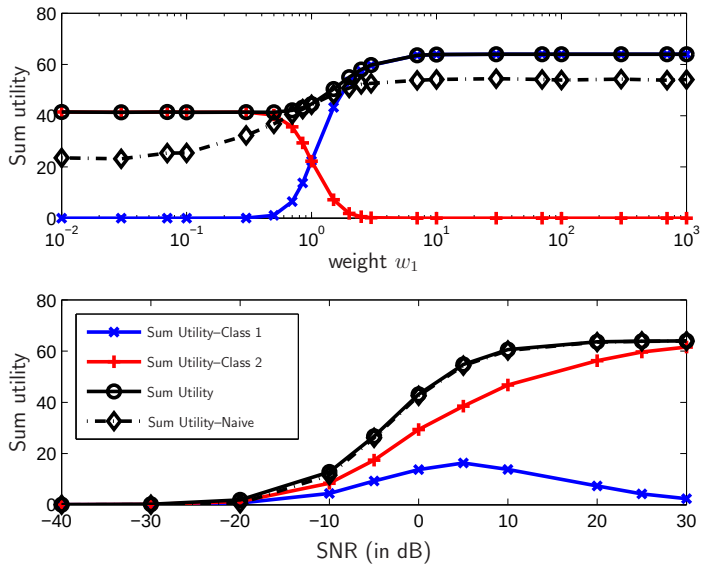

Figure 3.7: The top plot shows sum utility versus w_1 when $w_2 = 1$, $\mathsf{SNR} = 0$ dB. The bottom plot shows the sum-utility versus SNR when $w_1 = 0.85, w_2 = 1$. Here, $N = 64$, $K = 16$, and $\mathsf{SNR}_{\mathsf{pilot}} = -10$ dB.

In Figure 3.8, we compare the performances of our proposed algorithms to the state-of-the-art algorithms in [31, 33]. In particular, we first compare the golden-section-search based algorithm from [31] to our CSRA algorithm. For CSRA, we choose the utility function and the SNR distributions to maximize the upper bound on capacity computed via the effective SNR $\frac{1}{K}\sum_{n,k}\log\left(1+\frac{p_{n,k,1}|\hat{h}_{n,k}|^4}{|\hat{h}_{n,k}|^2+\sigma_e^2\, p_{n,k,1}|\hat{h}_{n,k}|^2}\right)$ from [31, Eq. (4)]. Second, we compare the subgradient-based algorithm proposed for discrete allocation in [33] to our DSRA algorithm. For DSRA, we choose the utility $U_{n,k,m}(g) = \frac{1}{K}\log(1-\log(1-g))\ \forall n,k,m$, so that we maximize $\frac{1}{K}\sum_{n,k}\mathrm{E}\{\log(1+p_{n,k,1}\gamma_{n,k})\}$, as

in [33]. The top plot in Fig. 3.8 shows the mean deviation of the estimated value of the dual variable μ from the optimum (i.e., μ^*), and the bottom plot shows the total utility achieved as a function of the number of μ-updates. For the subgradient-based algorithm in [33], we set the step-size in the i^{th} μ-update to be $1/i$. In the top plot, it can be seen that the proposed algorithms outperform the algorithms in [31,33] and converge toward μ^* at a much faster rate. The bottom plot shows that the proposed algorithms achieve a much higher utility than the algorithms in [31, 33] for the first few μ-updates, illustrating the speed of our approaches. Note that the golden-section algorithm only provides estimates of μ^* at even numbers of μ-updates.

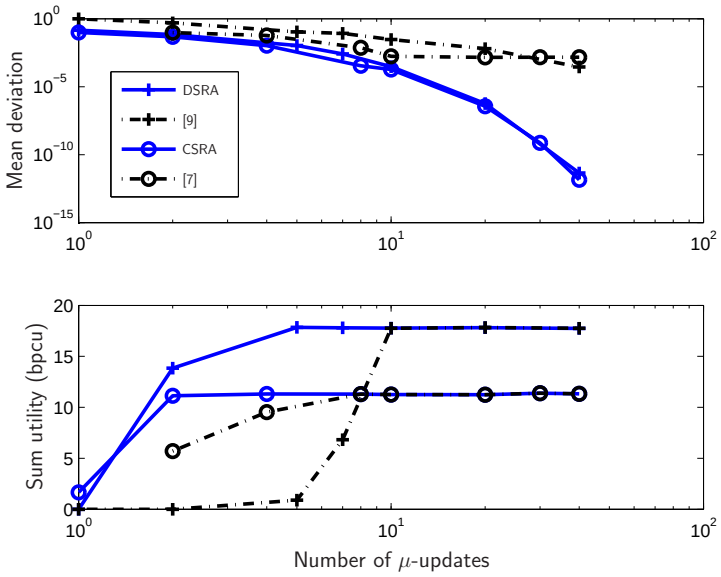

Figure 3.8: The top plot shows the mean deviation of the estimated dual variable μ from μ^*, and the bottom plot shows average sum-utility, as a function of the number of μ-updates. Here, $N = 64$, $K = 16$, $\mathsf{SNR} = 10$ dB, and $\mathsf{SNR}_{\mathsf{pilot}} = -10$ dB.

3.7 Conclusion

In this chapter, we considered the problem of joint scheduling and resource allocation (SRA) in downlink OFDMA systems under imperfect channel-state information. We considered two scenarios: 1) when subchannel sharing is allowed, and 2) when it is not. Both cases were framed as optimization problems that maximize a utility function subject to a sum-power constraint. Although the optimization problem in the first scenario (the so-called "continuous" or CSRA case) was found to be non-convex, we showed that it can be converted to a convex optimization problem and solved using a dual optimization approach with zero duality gap. An algorithmic implementation of the CSRA solution was also provided. The optimization problem faced in the second scenario (the so-called "discrete" or DSRA case) was found to be a mixed-integer programming problem. To attack it, we linked the DSRA problem to the CSRA problem, and showed that, in some cases, the DSRA solution coincides with the CSRA solution. For the case that the solutions do not coincide, we proposed a practical DSRA algorithm and bounded its performance. Numerical results were then presented under a variety of settings. The performance of the proposed CSRA and DSRA algorithms schemes under imperfect CSI were compared to those under perfect CSI and no instantaneous CSI (i.e., fixed-power random scheduling). In all cases, it was found that the proposed imperfect-CSI-based algorithms offer a significant advantage over schemes that do not use instantaneous CSI. Next, our DSRA bound was numerically evaluated and found to be extremely tight. We then demonstrated an application of DSRA to maximization of a pricing-based utility. Finally, our CSRA and DSRA algorithms were compared to the state-of-the-art

golden-section-search [31] and subgradient [33] based algorithms and shown to yield significant improvements in convergence rate.

Table 3.1: Algorithmic implementations of the proposed algorithms

Proposed CSRA algorithm

1. Set $\underline{\mu} = \mu_{\mathsf{min}}$, $\bar{\mu} = \mu_{\mathsf{max}}$, and $\mu = \frac{\underline{\mu}+\bar{\mu}}{2}$.
2. For each subchannel $n = 1, \ldots, N$:
   (a) For each (k, m),
      i. Use (3.11) and (3.13) to calculate $p^*_{n,k,m}(\mu)$.
      ii. Use (3.14) to calculate $V_{n,k,m}(\mu, p^*_{n,k,m}(\mu))$.
   (b) Calculate $S_n(\mu)$ using (3.15).
3. If $|S_n(\mu)| \leq 1 \; \forall n$, then find $\mathbf{I}^*(\mu)$ using (3.16), else use (3.30) and set $\mathbf{I}^*(\mu) = \mathbf{I}^{\mathsf{min}}(\mu)$.
4. Find $\mathbf{x}^*(\mu, \mathbf{I}^*(\mu))$ using (3.12) and calculate $X^*_{\mathsf{tot}}(\mu) = \sum_{n,k,m} x^*_{n,k,m}(\mu, \mathbf{I}^*(\mu))$.
5. If $X^*_{\mathsf{tot}}(\mu) \geq P_{\mathsf{con}}$, set $\underline{\mu} = \mu$, otherwise set $\bar{\mu} = \mu$.
6. If $\bar{\mu} - \underline{\mu} > \kappa$, go to step 2), else proceed.
7. Now we have $\mu^* \in [\underline{\mu}, \bar{\mu}]$ and $\bar{\mu} - \underline{\mu} < \kappa$. If $X^*_{\mathsf{tot}}(\underline{\mu}) \neq X^*_{\mathsf{tot}}(\bar{\mu})$, set $\lambda = \frac{X^*_{\mathsf{tot}}(\underline{\mu}) - P_{\mathsf{con}}}{X^*_{\mathsf{tot}}(\underline{\mu}) - X^*_{\mathsf{tot}}(\bar{\mu})}$, else set $\lambda = 0$.
8. The optimal user-MCS allocation is given by $\hat{\mathbf{I}}_{\mathsf{CSRA}} = \lambda \mathbf{I}^*(\bar{\mu}) + (1-\lambda)\mathbf{I}^*(\underline{\mu})$ and the corresponding optimal $\mathbf{x}$ is given by $\hat{\mathbf{x}}_{\mathsf{CSRA}} = \lambda \mathbf{x}^*(\bar{\mu}, \mathbf{I}^*(\bar{\mu})) + (1-\lambda)\mathbf{x}^*(\underline{\mu}, \mathbf{I}^*(\underline{\mu}))$. The optimal power allocation, $\hat{\mathbf{p}}_{\mathsf{CSRA}}$, then can be found using

$$\hat{p}_{n,k,m,\mathsf{CSRA}} = \begin{cases} \frac{\hat{x}_{n,k,m,\mathsf{CSRA}}}{\hat{I}_{n,k,m,\mathsf{CSRA}}} & \text{if } \hat{I}_{n,k,m,\mathsf{CSRA}} \neq 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $\hat{I}_{n,k,m,\mathsf{CSRA}}$ and $\hat{x}_{n,k,m,\mathsf{CSRA}}$ denote the $(n,k,m)^{\text{th}}$ component of $\hat{\mathbf{I}}_{\mathsf{CSRA}}$ and $\hat{\mathbf{x}}_{\mathsf{CSRA}}$, respectively. Notice that the obtained solution satisfies the sum-power constraint with equality.

Brute force algorithm for a given $\mathbf{I}$

1. Initialize $\underline{\mu} = \mu_{\mathsf{min}}$ and $\bar{\mu} = \mu_{\mathsf{max}}$.
2. Set $\mu = \frac{\underline{\mu}+\bar{\mu}}{2}$.
3. For each (n, k, m), use (3.38)-(3.40) to obtain $x^*_{n,k,m}(\mu)$.
4. Find $X^*_{\mathsf{tot}}(\mathbf{I}, \mu)$ using (3.41).
5. If $X^*_{\mathsf{tot}}(\mathbf{I}, \mu) > P_{\mathsf{con}}$, set $\underline{\mu} = \mu$, otherwise set $\bar{\mu} = \mu$.
6. If $\bar{\mu} - \underline{\mu} < \kappa$, go to step 7), otherwise go to step 2).
7. If $X^*_{\mathsf{tot}}(\mathbf{I}, \bar{\mu}) \neq X^*_{\mathsf{tot}}(\mathbf{I}, \underline{\mu})$, set $\lambda = \frac{X^*_{\mathsf{tot}}(\mathbf{I}, \underline{\mu}) - P_{\mathsf{con}}}{X^*_{\mathsf{tot}}(\mathbf{I}, \underline{\mu}) - X^*_{\mathsf{tot}}(\mathbf{I}, \bar{\mu})}$, otherwise set $\lambda = 0$.
8. Set $\hat{\mu}_{\mathbf{I}} = \bar{\mu}$. The best actual power allocation is given by $\hat{\mathbf{x}}_{\mathbf{I}} = \lambda \mathbf{x}^*(\bar{\mu}) + (1-\lambda)\mathbf{x}^*(\underline{\mu})$ and the best power allocation, $\hat{\mathbf{p}}_{\mathbf{I}}$, is given by

$$\hat{p}_{n,k,m,\mathbf{I}} = \begin{cases} \frac{\hat{x}_{n,k,m,\mathbf{I}}}{I_{n,k,m}} & \text{if } I_{n,k,m} \neq 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $\hat{p}_{n,k,m,\mathbf{I}}$ and $\hat{x}_{n,k,m,\mathbf{I}}$ are the $(n,k,m)^{\text{th}}$ element of $\hat{\mathbf{p}}_{\mathbf{I}}$ and $\hat{\mathbf{x}}_{\mathbf{I}}$, respectively. The corresponding Lagrangian, found using $\hat{L}_{\mathbf{I}} = L_{\mathbf{I}}(\bar{\mu}, \mathbf{p}^*(\mu))$, gives the optimal Lagrangian value.

Proposed DSRA algorithm

1. Use the algorithmic implementation of the proposed CSRA solution in to find $\mathbf{I}^*(\underline{\mu})$ and $\mathbf{I}^*(\bar{\mu})$, where the optimal μ for the CSRA problem, i.e., μ^* lies in the set $[\underline{\mu}, \bar{\mu}]$, $\bar{\mu} - \underline{\mu} < \kappa$, and $\mathbf{I}^*(\underline{\mu}), \mathbf{I}^*(\bar{\mu}) \in \mathcal{I}_{\mathsf{DSRA}}$.
2. For both $\mathbf{I} = \mathbf{I}^*(\underline{\mu})$ and $\mathbf{I} = \mathbf{I}^*(\bar{\mu})$ (since they may differ), calculate $\hat{\mathbf{p}}_{\mathbf{I}}$ and $\hat{L}_{\mathbf{I}}$ as described for the brute force algorithm.
3. Choose $\hat{\mathbf{I}}_{\mathsf{DSRA}} = \text{argmin}_{\mathbf{I} \in \{\mathbf{I}^*(\underline{\mu}), \mathbf{I}^*(\bar{\mu})\}} \hat{L}_{\mathbf{I}}$ as the user-MCS allocation and $\hat{\mathbf{p}}_{\mathsf{DSRA}} = \hat{\mathbf{p}}_{\hat{\mathbf{I}}_{\mathsf{DSRA}}}$ as the associated power allocation.

Chapter 4: Joint Scheduling and Resource Allocation in OFDMA Downlink Systems via ACK/NAK Feedback

4.1 Introduction

The OFDMA scheduling-and-resource-allocation problem has been addressed in a number of studies that assume the availability of perfect channel state information (CSI) at the BS (e.g., [34–37, 39, 44, 54]). As mentioned in the previous chapter, in practice, it is difficult for the BS to maintain perfect CSI for all users and all subchannels (particularly when the number of users is large) since CSI is most easily obtained at the user terminals, and the bandwidth available for feedback of CSI to the BS is scarce. Hence, practical resource allocation schemes use some form of limited feedback [55], such as quantized channel gains.

In the previous chapter, we studied an instantaneous resource allocation problem under the availability of a generic distribution on the channnel-gains of users, and presented results assuming pilot-aided channel knowledge at the base-station. In this chapter, we consider the exclusive use of ACK/NAK feedback, as provided by the automatic repeat request (ARQ) [56] mechanism present in most wireless downlinks. We assume standard ARQ,[14] where every scheduled user provides the BS with either

[14] The approach we develop in this chapter could be easily extended to other forms of link-layer feedback, e.g., Type-I and Type-II Hybrid ARQ. For simplicity and ease of exposition, however, we consider only standard ARQ.

an acknowledgment (ACK), if the most recent data packet has been correctly decoded, or a negative acknowledgment (NAK), if not. Although ACK/NAKs do not provide direct information about the state of the channel, they do provide *relative* information about channel quality that can be used for the purpose of transmitter adaptation (e.g., [57, 58]). For example, if an NAK was received for a particular packet, then it is likely that the subchannel's signal-to-noise ratio (SNR) was below that required to support the transmission rate used for that packet. We consider the *exclusive* use ACK/NAK feedback provided by the link layer, because this allows us to completely avoid *any additional* feedback, such as feedback about quantized channel gains.

There are interesting implications to the use of (quantized) *error-rate* feedback (like ACK/NAK) for transmitter adaptation, as opposed to quantized channel-state feedback. With error-rate feedback, the transmission parameters applied at a given time-slot affect not only the throughput for that slot, but also the corresponding feedback, which will impact the quality of future transmitter-CSI, and thus future throughput. For example, if the transmission parameters are chosen to maximize only the instantaneous throughput, e.g., by scheduling those users that the BS believes are currently best, then little will be learned about the changing states of other user channels, implying that future scheduling decisions may be compromised. On the other hand, if the BS schedules not-recently-scheduled users solely for the purpose of probing their channels, then instantaneous throughput may be compromised. Thus, when using error-rate feedback, the BS must navigate the classic tradeoff between exploitation and exploration [59].

In this work, we propose a scheme whereby the BS uses ACK/NAK feedback to maintain a posterior channel distribution for every user and, from these distributions, performs simultaneous user subchannel-scheduling, power-allocation, and rate-selection. In doing so, the BS aims to maximize an expected, long-term, generic *utility* criterion that is a function of the per-user/channel/rate goodputs. Our use of a generic utility-based criterion allows us to handle, e.g., sum-capacity maximization, throughput maximization under practical modulation-and-coding schemes, and throughput-based pricing (e.g., [26–28]), as discussed in the sequel. To this end, we exploit the results in Chapter 3, which offers an efficient near-optimal scheme for utility-based OFDMA resource allocation under probabilistic CSI. Our use of ACK/NAK-feedback, however, makes our problem considerably more complicated than the one considered in Chapter 3. For example, as we show in the sequel, the optimal solution to our expected long-term utility-maximization problem is a *partially observable Markov decision process* (POMDP) that would involve the solution of many mixed-integer optimization problems during each time-slot. Due to the impracticality of the POMDP solution, we instead consider (suboptimal) *greedy* utility-maximization schemes. As justification for this approach, we first establish that the optimal utility maximization strategy would itself be greedy if the BS had perfect CSI for all user-subchannel combinations. Moreover, we establish that the performance of this perfect-CSI (greedy) scheme upper-bounds the optimal ACK/NAK-feedback-based (POMDP) scheme. We then propose a novel, greedy utility-maximization scheme whose performance is shown (via the upper bound) to be close to optimal. Finally, due to the computational demands of tracking the posterior channel distribution for every user, we propose a low-complexity implementation based on particle filtering.

The rest of the chapter is organized as follows. In Section 4.2, we discuss some past related works. In Section 4.3, we outline the system model and, in Section 4.4, we investigate the optimal scheduling and resource allocation scheme. Due to the implementation complexity of the optimal scheme, we propose a suboptimal greedy scheme in Section 4.5 that maintains posterior channel distributions inferred from the received ACK/NAK feedback. In Section 4.6, we show how these posteriors can be recursively updated via particle filtering. Numerical results are presented in Section 4.7, and conclusions are stated in Section 4.8.

4.2 Past Work

We now describe the relation of our work to the existing literature [60–62]. In [60], a *learning-automata*-based user/rate scheduling algorithm was proposed to maximize system throughput based on ACK/NAK feedback while satisfying per-user throughput constraints. While [60] considered a single channel, we consider joint user/rate scheduling and power allocation in a multi-channel OFDMA setting. In [61], a state-space-based approach was taken to jointly schedule users/rates and allocate powers in downlink OFDMA systems under slow-fading channels in the presence of ACK/NAK feedback and imperfect subchannel-gain estimates at the BS. In particular, assuming a discrete channel model, goodput maximization was considered under a target maximum packet-error probability constraint and a sum-power constraint across all time-slots. Its solution led to a POMDP which was solved using a dynamic-program. While the approach in [61] is applicable to only goodput maximization under discrete-state channels, ours is applicable to generic utility maximization problems under continuous-state channels. Furthermore, our approach is based on particle

filtering and lends itself to practical implementation. In [62], the user/rate scheduling and power allocation problem in OFDMA systems with quasi-static channels and ACK/NAK feedback was formulated as a Markov Decision Process and an efficient algorithm was proposed to maximize achievable sum-rate while maintaining a target packet-error-rate and a sum-power constraint over a finite time-horizon. Apart from assuming a discrete-state quasi-static channel model, the scope of this work was limited by two other assumptions: *i)* in each time-slot, the BS scheduled only one user across all subchannels for data transmission, and *ii)* all users decoded the broadcasted data-packet and sent ACK/NAK feedback to the BS. In contrast, we consider the scenario where multi-user diversity is efficiently exploited by scheduling different users across different subchannels, and only the scheduled users report ACK/NAK feedback. Furthermore, we consider general utility maximization under continuous-state time-varying channels, and propose a polynomial-complexity joint scheduling and resource allocation scheme with provable performance guarantees.

4.3 System Model

We consider a packetized downlink OFDMA system with a pool of K users. During each time slot, the BS (i.e., "controller") transmits packets of data, composed of codewords from a generic signaling scheme, through N OFDMA subchannels (with $N \leqslant K$). Each packet propagates through a fading channel on the way to its intended mobile user, where the fading channel is assumed to be time-invariant over the packet duration, but is allowed to vary across packets in a Markovian manner. Henceforth, we will use "time" when referring to the packet index. At each time-instant, the BS

must decide—for each subchannel—which user to schedule, which modulation-and-coding scheme (MCS) to use, and how much power to allocate.

We assume M choices of MCS, where the MCS index $m \in \{1, \ldots, M\}$ corresponds to a transmission rate of r_m bits per packet and a packet error rate of the form $\epsilon = a_m e^{-b_m P \gamma}$ under transmit power P and squared subchannel gain (SSG) γ, where a_m and b_m are constants [33]. Let (n, k, m) represent the combination of user k and MCS m over subchannel n. In the sequel, we use $P^t_{n,k,m}$, $\gamma^t_{n,k}$, and $\epsilon^t_{n,k,m}$ to denote—respectively—the power allocated to, the SSG experienced by, and the error rate of the combination (n, k, m) at time t. Additionally, we denote the scheduling decision by $I^t_{n,k,m} \in \{0, 1\}$, where $I^t_{n,k,m} = 1$ indicates that user/rate (k, m) was scheduled on subchannel n at time t, whereas $I^t_{n,k,m} = 0$ indicates otherwise. Since we assume that only one user/rate (k, m) can be scheduled on a given subchannel n at a given time t, we have the "subchannel resource" constraint $\sum_{k,m} I^t_{n,k,m} \leqslant 1$ for all n, t. We also assume a "sum-power constraint" of the form $\sum_{n,k,m} I^t_{n,k,m} P^t_{n,k,m} \leqslant X_{\mathsf{con}}$ for all t.

Our goal in scheduling and resource allocation is to maximize an expected long-term utility criterion that is a function of the per-user/rate/subchannel goodputs, i.e., $\mathrm{E}\left\{ \sum_{n,k,m,t} U_{n,k,m}(g^t_{n,k,m}) \right\}$. Here, $g^t_{n,k,m}$ denotes the goodput contributed by user k with MCS m on subchannel n at time t, which can be expanded as $g^t_{n,k,m} = I^t_{n,k,m}(1 - \epsilon^t_{n,k,m}) r_m$. Meanwhile, $U_{n,k,m}(\cdot)$ is a generic utility function that we assume (for technical reasons) is twice differentiable, strictly-increasing, and concave, with $U_{n,k,m}(0) < \infty$. We use $U_{n,k,m}(\cdot)$ to transform goodput into other metrics that are more meaningful from the perspective of quality-of-service (QoS), fairness [63], or pricing (e.g., [26–28]). For example, to maximize sum-goodput, one would simply use $U_{n,k,m}(x) = x$. To enforce fairness across users, one could instead maximize

weighted sum-goodput via $U_{n,k,m}(x) = w_k x$, where $\{w_k\}$ are appropriately chosen user-dependent weights. To maximize sum capacity, i.e., $\sum_{n,k} I^t_{n,k,1} \log(1+P^t_{n,k,1}\gamma^t_{n,k})$, one would choose $M = a_1 = b_1 = r_1 = 1$ and $U_{n,k,1}(x) = \log(1 - \log(1 - x))$ for $x \in [0, 1)$. To incorporate user-fairness into capacity maximization, one could instead choose $U_{n,k,1}(\cdot) = w_k \log(1 - \log(1 - x))$, where again $\{w_k\}$ are appropriately chosen user-dependent weights [33].

For each time t, the BS performs scheduling and resource allocation based on posterior distributions on the SSGs $\{\gamma^t_{n,k}\}$ inferred from previously received ACK/NAK feedback. In the sequel, we write the ACK/NAK feedback about the packet transmitted to user k across subchannel n at time t by $f^t_{n,k} \in \{1, 0, \emptyset\}$, where 1 indicates an ACK, 0 indicates a NAK, and $\emptyset$ covers the case that user k was not scheduled on subchannel n at time t. Thus, in the case of an infinite past horizon and a feedback delay of $d \geqslant 1$ packets, the BS would have access to the feedbacks $\{f^\tau_{n,k}\ \forall n, k\}^{t-d}_{\tau=-\infty}$ for time-t scheduling.

4.4 Optimal Scheduling and Resource Allocation

In this section, we describe the optimal solution to the problem of scheduling and resource allocation over the finite time-horizon $t \in \{1, \ldots, T\}$. For this purpose, some additional notation will be useful. To denote the collection of all time-t scheduling variables $\{I^t_{n,k,m}\}$, we use $\mathbf{I}^t \in \{0, 1\}^{NKM}$. To denote the collection of all time-t powers $\{P^t_{n,k,m}\}$, we use $\mathbf{P}^t \in [0, \infty)^{NKM}$. To denote the collection of all time-t ACK/NAK feedbacks $\{f^t_{n,k}\}$ we use $\mathbf{F}^t \in \{1, 0, \emptyset\}^{NK}$, and to denote the collection of all time-t user-k feedbacks we use $\mathbf{f}^t_k \in \{1, 0, \emptyset\}^N$.

For time-t scheduling and resource allocation, the controller has access to the previous feedback $\mathbf{F}_{-\infty}^{t-d} \triangleq \{\mathbf{F}^{-\infty}, \ldots, \mathbf{F}^{t-d}\}$, scheduling decisions $\mathbf{I}_{-\infty}^{t-d} \triangleq \{\mathbf{I}^{-\infty}, \ldots, \mathbf{I}^{t-d}\}$, and power allocations $\mathbf{P}_{-\infty}^{t-d} \triangleq \{\mathbf{P}^{-\infty}, \ldots, \mathbf{P}^{t-d}\}$. It then uses this knowledge to determine the schedule $\mathbf{I}^t$ and power allocation $\mathbf{p}^t$ maximizing the expected utility of the current and remaining packets:

$$\begin{aligned}
&(\mathbf{I}^{t,\mathsf{opt}}, \mathbf{P}^{t,\mathsf{opt}}) \\
&\quad = \underset{(\mathbf{I}^t,\mathbf{P}^t)\in\mathcal{X}}{\operatorname{argmax}} \, \mathrm{E}\Bigg\{ \sum_{n,k,m} I_{n,k,m}^t U_{n,k,m}\big((1 - a_m e^{-b_m P_{n,k,m}^t \gamma_{n,k}^t}) r_m\big) \\
&\qquad + \sum_{\tau=t+1}^{T} I_{n,k,m}^{\tau,\mathsf{opt}} U_{n,k,m}\big((1 - a_m e^{-b_m P_{n,k,m}^{\tau,\mathsf{opt}} \gamma_{n,k}^{\tau}}) r_m\big) \Bigg| \mathbf{F}_{-\infty}^{t-d}, \mathbf{I}_{-\infty}^{t-d}, \mathbf{P}_{-\infty}^{t-d} \Bigg\}, \quad (4.1)
\end{aligned}$$

where the domain of $\mathbf{I}^t$ is $\mathcal{I} \triangleq \{\mathbf{I} \in \{0,1\}^{NKM} : \sum_{k,m} I_{n,k,m} \leqslant 1 \ \forall n\}$, the domain of $\mathbf{P}^t$ is $\mathcal{P} \triangleq [0,\infty)^{NKM}$, and $\mathcal{X} \triangleq \{(\mathbf{I},\mathbf{P}) \in \mathcal{I} \times \mathcal{P} : \sum_{n,k,m} I_{n,k,m} P_{n,k,m} \leqslant X_{\mathsf{con}}\}$. The expectation in (4.1) is jointly over the squared subchannel gains (SSGs) $\{\gamma_{n,k}^{\tau} : \tau = t, \ldots, T, \forall n, \forall k\}$. Using the abbreviations $\tilde{U}_{n,k,m}^t(I_{n,k,m}, P_{n,k,m}) \triangleq I_{n,k,m} U_{n,k,m}\big((1 - a_m e^{-b_m P_{n,k,m} \gamma_{n,k}^t}) r_m\big)$ and $\mathbb{F}_{-\infty}^{t-d} \triangleq \{\mathbf{F}_{-\infty}^{t-d}, \mathbf{I}_{-\infty}^{t-d}, \mathbf{P}_{-\infty}^{t-d}\}$, the optimal expected utility over the remaining packets $\{t, \ldots, T\}$ can be written (for $t \geqslant 0$) as

$$U_{\mathsf{tot}}^{t,\mathsf{opt}}(\mathbb{F}_{-\infty}^{t-d}) \triangleq \mathrm{E}\Bigg\{ \sum_{\tau=t}^{T} \sum_{n,k,m} \tilde{U}_{n,k,m}^{\tau}\big(I_{n,k,m}^{\tau,\mathsf{opt}}, P_{n,k,m}^{\tau,\mathsf{opt}}\big) \,\Bigg|\, \mathbb{F}_{-\infty}^{t-d} \Bigg\}. \quad (4.2)$$

For a unit-delay[15] system (i.e. $d = 1$), the following Bellman equation [49] specifies the corresponding finite-horizon dynamic program:

$$\begin{aligned}
U_{\mathsf{tot}}^{t,\mathsf{opt}}(\mathbb{F}_{-\infty}^{t-1}) \;=\; & \max_{(\mathbf{I}^t,\mathbf{P}^t)\in\mathcal{X}} \Bigg[\mathrm{E}\Bigg\{ \sum_{n,k,m} \tilde{U}_{n,k,m}^t\big(I_{n,k,m}^t, P_{n,k,m}^t\big) \,\Bigg|\, \mathbb{F}_{-\infty}^{t-1} \Bigg\} \\
& + \mathrm{E}\Big\{ U_{\mathsf{tot}}^{t+1,\mathsf{opt}}\big(\mathbb{F}_{-\infty}^{t-1} \cup \{\mathbf{F}^t, \mathbf{I}^t, \mathbf{P}^t\}\big) \,\Big|\, \mathbb{F}_{-\infty}^{t-1} \Big\} \Bigg], \quad (4.3)
\end{aligned}$$

[15] For the $d > 1$ case, the Bellman equation is more complicated, and so we omit it for brevity.

where the second expectation is over the feedbacks $\mathbf{F}^t$. The solution obtained by solving (4.1) is typically referred to as a partially observable Markov decision process (POMDP) [59].

The definition of $\mathcal{P}$ implies that the controller has an uncountably infinite number of possible actions. Although this could be circumvented (at the expense of performance) by restricting the powers $P^t_{n,k,m}$ to come from a finite set, the problem would remain very complex due to the continuous-state nature of the SSGs $\gamma^t_{n,k}$. While these SSGs could then be quantized (causing additional performance loss), the problem would still remain computationally intensive, since POMDPs (even with finite states and actions) are PSPACE-complete, i.e., they require both complexity and memory that grow exponentially with the horizon T [64]. To see why, notice from (4.3) that the solution of the problem at every time t depends on the optimal solution at times up to $t-1$. Because both terms on the right side of (4.3) are dependent on $(\mathbf{I}^t, \mathbf{P}^t)$, however, the solution of the problem at time t also depends on the solution of the problem at time $t+1$, which in turn depends on the solution of the problem at time $t+2$, and so on. In conclusion, the optimal controller is not practical to implement, even under power/SSG quantization.

Consequently, we will turn our attention to (sub-optimal) *greedy* strategies, i.e., those that do not consider the effect of current actions on future utilities. To better understand their performance relative to that of the optimal POMDP, we derive an upper bound on POMDP performance.

4.4.1 The "Causal Global Genie" Upper Bound

Our POMDP-performance upper-bound, which we will refer to as the "causal global genie" (CGG), is based on the presumption of *perfect* error-rate feedback of *all* previous user/subchannel combinations, i.e., $\{\epsilon^{\tau}_{n,k,m}\,\forall n,k,\tau \leqslant t-d\}$. For comparison, the ACK/NAK feedback available to the POMDP is a form of *degraded* error-rate feedback on *previously scheduled* user/subchannel combinations. Since, given knowledge of $\epsilon^{\tau}_{n,k,m}$ and $P^{\tau}_{n,k,m}$ for any rate index m, the SSG $\gamma^{\tau}_{n,k}$ can be obtained by simply inverting the error-rate expression $\epsilon^{\tau}_{n,k,m} = a_m e^{-b_m P^{\tau}_{n,k,m}\gamma^{\tau}_{n,k}}$, our genie-aided bound is based, equivalently, on perfect feedback of all previous SSGs $\{\gamma^{\tau}_{n,k}\,\forall n,k,\tau \leqslant t-d\}$. In the sequel, we use $\gamma^t \in [0,\infty)^{NK}$ to denote the collection of all time-t SSGs $\{\gamma^t_{n,k}\,\forall k,n\}$, and we define $\gamma^{t-d}_{-\infty} \triangleq \{\gamma^{-\infty},\ldots,\gamma^{t-d}\}$.

We characterize the CGG as "global" since it uses feedback from *all* user/subchannel combinations, not just the previously scheduled ones. Although a tighter bound might result if the (perfect) error-rate feedback was restricted to only previously scheduled user/subchannel pairs, the bounding solution would remain a POMDP with an uncountable number of state-action pairs, making it impractical to evaluate. Evaluating the performance of the CGG, however, is straightforward since—under CGG feedback—optimal scheduling and resource maximization can be performed greedily. To see why, notice that, for any scheduling time $t \geqslant 0$, the CGG scheme allocates resources according to the following mixed-integer optimization problem:

$$\begin{aligned}(\mathbf{I}^{t,\mathsf{cgg}},\mathbf{P}^{t,\mathsf{cgg}}) \;=\; &\underset{(\mathbf{I}^t,\mathbf{P}^t)\in\mathcal{X}}{\operatorname{argmax}} \sum_{n,k,m} \mathrm{E}\Bigg\{\tilde{U}^t_{n,k,m}\big(I^t_{n,k,m},P^t_{n,k,m}\big) \\ &+ \sum_{\tau=t+1}^{T} \tilde{U}^{\tau}_{n,k,m}\big(I^{\tau,\mathsf{cgg}}_{n,k,m},P^{\tau,\mathsf{cgg}}_{n,k,m}\big)\Big|\mathbf{I}^{t-d}_{-\infty},\mathbf{P}^{t-d}_{-\infty},\gamma^{t-d}_{-\infty}\Bigg\}. \qquad (4.4)\end{aligned}$$

Since the choice of $\big\{(\mathbf{I}^{t+1,\mathsf{cgg}},\mathbf{P}^{t+1,\mathsf{cgg}}),\ldots,(\mathbf{I}^{T,\mathsf{cgg}},\mathbf{P}^{T,\mathsf{cgg}})\big\}$ does not depend on the choice of $(\mathbf{I}^{t,\mathsf{cgg}},\mathbf{P}^{t,\mathsf{cgg}})$, the previous optimization problem simplifies to

$$(\mathbf{I}^{t,\mathsf{cgg}},\mathbf{P}^{t,\mathsf{cgg}}) \;=\; \operatorname*{argmax}_{(\mathbf{I}^t,\mathbf{P}^t)\in\mathcal{X}} \sum_{n,k,m} \mathrm{E}\left\{\tilde{U}^t_{n,k,m}\big(I^t_{n,k,m},P^t_{n,k,m}\big)\Big|\gamma^{t-d}_{-\infty}\right\}. \tag{4.5}$$

In the following lemma, we formally establish that the utility achieved by the CGG upper-bounds that achieved by the optimal POMDP controller with ACK/NAK feedback.

Lemma 7. *Given arbitrary past allocations $(\mathbf{I}^{t-d}_{-\infty},\mathbf{P}^{t-d}_{-\infty})$, and the corresponding ACK/NAKs $\mathbf{F}^{t-d}_{-\infty}$, the expected total utility for optimal resource allocation under the latter feedback is no higher than the expected total utility under CGG feedback, i.e.,*

$$\begin{aligned}&\sum_{n,k,m}\sum_{\tau=t}^{T}\mathrm{E}\left\{\tilde{U}^\tau_{n,k,m}\big(I^{\tau,\mathsf{opt}}_{n,k,m},P^{\tau,\mathsf{opt}}_{n,k,m}\big)\;\Big|\;\mathbb{F}^{t-d}_{-\infty}\right\}\\ &\quad\leqslant\;\sum_{n,k,m}\sum_{\tau=t}^{T}\mathrm{E}\left\{\tilde{U}^\tau_{n,k,m}\big(I^{\tau,\mathsf{cgg}}_{n,k,m},P^{\tau,\mathsf{cgg}}_{n,k,m}\big)\;\Big|\;\mathbb{F}^{t-d}_{-\infty}\right\}.\end{aligned} \tag{4.6}$$

Proof. For any $\tau\in\{t,\ldots,T\}$ and any realization of $(\mathbf{I}^{t-d}_{-\infty},\mathbf{P}^{t-d}_{-\infty})$, we can write

$$\begin{aligned}&\mathrm{E}\Big\{\sum_{n,k,m}\tilde{U}^\tau_{n,k,m}\big(I^{\tau,\mathsf{opt}}_{n,k,m},P^{\tau,\mathsf{opt}}_{n,k,m}\big)\big|\mathbb{F}^{t-d}_{-\infty}\Big\}\\
&\quad\leqslant\;\operatorname*{argmax}_{(\mathbf{I}^\tau,\mathbf{P}^\tau)\in\mathcal{X}}\mathrm{E}\Big\{\sum_{n,k,m}\tilde{U}^\tau_{n,k,m}\big(I^{\tau}_{n,k,m},P^{\tau}_{n,k,m}\big)\;\big|\;\mathbb{F}^{t-d}_{-\infty}\Big\} &(4.7)\\
&\quad=\;\operatorname*{argmax}_{(\mathbf{I}^\tau,\mathbf{P}^\tau)\in\mathcal{X}}\mathrm{E}\Big\{\mathrm{E}\Big\{\sum_{n,k,m}\tilde{U}^\tau_{n,k,m}\big(I^{\tau}_{n,k,m},P^{\tau}_{n,k,m}\big)\;\big|\;\mathbb{F}^{t-d}_{-\infty},\gamma^{t-d}_{-\infty}\Big\}\;\Big|\;\mathbb{F}^{t-d}_{-\infty}\Big\} &(4.8)\\
&\quad\leqslant\;\mathrm{E}\Big\{\operatorname*{argmax}_{(\mathbf{I}^\tau,\mathbf{P}^\tau)\in\mathcal{X}}\mathrm{E}\Big\{\sum_{n,k,m}\tilde{U}^\tau_{n,k,m}\big(I^{\tau}_{n,k,m},P^{\tau}_{n,k,m}\big)\;\Big|\;\mathbb{F}^{t-d}_{-\infty},\gamma^{t-d}_{-\infty}\Big\}\;\Big|\;\mathbb{F}^{t-d}_{-\infty}\Big\} &(4.9)\\
&\quad=\;\mathrm{E}\Big\{\operatorname*{argmax}_{(\mathbf{I}^\tau,\mathbf{P}^\tau)\in\mathcal{X}}\mathrm{E}\Big\{\sum_{n,k,m}\tilde{U}^\tau_{n,k,m}\big(I^{\tau}_{n,k,m},P^{\tau}_{n,k,m}\big)\;\big|\;\mathbf{I}^{t-d}_{-\infty},\mathbf{P}^{t-d}_{-\infty},\gamma^{t-d}_{-\infty}\Big\}\;\Big|\;\mathbb{F}^{t-d}_{-\infty}\Big\} &(4.10)\\
&\quad=\;\mathrm{E}\Big\{\sum_{n,k,m}\tilde{U}^\tau_{n,k,m}\big(I^{\tau,\mathsf{cgg}}_{n,k,m},P^{\tau,\mathsf{cgg}}_{n,k,m}\big)\;\big|\;\mathbb{F}^{t-d}_{-\infty}\Big\}, &(4.11)\end{aligned}$$

where (4.7) follows since $(\mathbf{I}^{t,\mathsf{opt}},\mathbf{P}^{t,\mathsf{opt}})$ is chosen to maximize the long term sum-utility—not the instantaneous sum-utility; (4.9) follows since $\max_{y,z}\mathrm{E}\{f(y,z)\}\leqslant\mathrm{E}\{\max_{y,z}f(y,z)\}$ for any real-valued function $f(\cdot,\cdot)$; (4.10) follows by the definition of degraded feedback; and (4.11) follows by definition of the causal global genie. Finally, summing both sides of (4.11) over $\tau=\{t,\ldots,T\}$ yields (4.6). □

In the next section, we detail the greedy scheduling and resource allocation problem and propose a near-optimal solution.

4.5 Greedy Scheduling and Resource Allocation

The greedy scheduling and resource allocation (GSRA) problem is defined as follows.

$$\begin{aligned} \text{GSRA} \quad \triangleq \quad & \max_{\substack{\mathbf{I}^t \in \mathcal{I} \\ \mathbf{P}^t \in \mathcal{P}}} \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{m=1}^{M} I^t_{n,k,m} \, \mathrm{E}\left\{ U_{n,k,m}\big((1 - a_m e^{-b_m P^t_{n,k,m} \gamma^t_{n,k}}) r_m\big) \,\middle|\, \mathbb{F}^{t-d}_{-\infty} \right\} \\ & \text{s.t.} \quad \sum_{n,k,m} I^t_{n,k,m} \, P^t_{n,k,m} \leqslant X_{\mathsf{con}}. \end{aligned} \tag{4.12}$$

Note that, in contrast to the T-horizon objective (4.1), the greedy objective (4.12) does not consider the effect of $(\mathbf{I}^t, \mathbf{P}^t)$ on future utility. As stated earlier, we allow $U_{n,k,m}(\cdot)$ to be any real-valued function that is twice differentiable, strictly-increasing, and concave, with $U_{n,k,m}(0) < \infty$. Therefore, $U'_{n,k,m}(\cdot) > 0$ and $U''_{n,k,m}(\cdot) \leqslant 0$, using $'$ to denote the derivative.

Since it involves both discrete ($\mathbf{I}^t$) and continuous ($\mathbf{P}^t$) optimization variables, the GSRA problem (4.12) is a mixed-integer optimization problem. Such problems are generally NP-hard, meaning that polynomial-complexity solutions do not exist. Thus, in Section 4.5.2, we propose a *near*-optimal algorithm for (4.12) with polynomial complexity. To better explain that scheme, we first describe, in Section 4.5.1, a "brute force" optimal solution whose complexity grows exponentially in N, the number of subchannels.

4.5.1 Brute-Force Algorithm

The brute-force approach considers all possibilities of $\mathbf{I}^t \in \mathcal{I}$, each with the corresponding optimal power allocation. Supposing that $\mathbf{I}^t = \mathbf{I}$, the optimal power

allocation can be found by solving the convex optimization problem

$$\begin{aligned}
&\max_{\mathbf{P}\in\mathcal{P}} \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M} I_{n,k,m}\,\mathrm{E}\left\{U_{n,k,m}\big((1-a_m e^{-b_m P_{n,k,m}\gamma_{n,k}^t})r_m\big)\Big|\mathbb{F}_{-\infty}^{t-d}\right\} \\
&\text{s.t.} \quad \sum_{n,k,m} I_{n,k,m}\,P_{n,k,m} \leqslant X_{\mathsf{con}}. \qquad (4.13)
\end{aligned}$$

To proceed, we identify the Lagrangian associated with (4.13) as

$$\begin{aligned}
L_{\mathbf{I}}^t(\mu,\mathbf{P}) &= \Big(\sum_{n,k,m} I_{n,k,m}\,P_{n,k,m} - X_{\mathsf{con}}\Big)\mu \\
&\quad - \sum_{n,k,m} \mathrm{E}\left\{I_{n,k,m}U_{n,k,m}\big((1-a_m e^{-b_m P_{n,k,m}\gamma_{n,k}^t})r_m\big)\Big|\mathbb{F}_{-\infty}^{t-d}\right\}, \quad (4.14)
\end{aligned}$$

which yields the corresponding dual problem

$$\max_{\mu\geqslant 0}\min_{\mathbf{P}\in\mathcal{P}} L_{\mathbf{I}}^t(\mu,\mathbf{P}) = \max_{\mu\geqslant 0} L_{\mathbf{I}}^t(\mu,\mathbf{P}^*(\mu)) = L_{\mathbf{I}}^t(\mu_{\mathbf{I}}^*,\mathbf{P}^*(\mu_{\mathbf{I}}^*)), \qquad (4.15)$$

where $\mu_{\mathbf{I}}^*$ and $\mathbf{P}^*(\mu_{\mathbf{I}}^*)$ denote the optimal Lagrange multiplier and power allocation, respectively.

A detailed solution to (4.15) is given in Chapter 3, and so we describe only the main points here. First, for a given value of the Lagrange multiplier μ, it has been shown that the optimal powers equal

$$P_{n,k,m}^*(\mu) = \begin{cases} \tilde{P}_{n,k,m}(\mu) & \text{if } 0 \leqslant \mu \leqslant a_m b_m r_m U'_{n,k,m}\big((1-a_m)r_m\big)\,\mathrm{E}\left\{\gamma_{n,k}^t\big|\mathbb{F}_{-\infty}^{t-d}\right\} \\ 0 & \text{otherwise,} \end{cases} \qquad (4.16)$$

where $\tilde{P}_{n,k,m}(\mu)$ is defined as the (unique) solution to

$$\mu = a_m b_m r_m\,\mathrm{E}\left\{U'_{n,k,m}\big((1-a_m e^{-b_m \tilde{P}_{n,k,m}(\mu)\gamma_{n,k}^t})r_m\big)\gamma_{n,k}^t e^{-b_m \tilde{P}_{n,k,m}(\mu)\gamma_{n,k}^t}\Big|\mathbb{F}_{-\infty}^{t-d}\right\}. \qquad (4.17)$$

Then, for a given $\mathbf{I}$, the optimal value of μ (i.e., $\mu^*_{\mathbf{I}}$) obeys $\mu^*_{\mathbf{I}} \in [\mu_{\mathsf{min}}, \mu_{\mathsf{max}}] \subset (0, \infty)$, where

$$\mu_{\mathsf{min}} = \min_{n,k,m} a_m b_m r_m \, \mathrm{E}\left\{U'_{n,k,m}\big((1 - a_m e^{-b_m X_{\mathsf{con}} \gamma^t_{n,k}}) r_m\big) \gamma^t_{n,k} e^{-b_m X_{\mathsf{con}} \gamma^t_{n,k}} \big| \mathbb{F}^{t-d}_{-\infty}\right\}, \tag{4.18}$$

$$\mu_{\mathsf{max}} = \max_{n,k,m} a_m b_m r_m U'_{n,k,m}\big((1 - a_m) r_m\big) \, \mathrm{E}\left\{\gamma^t_{n,k} \big| \mathbb{F}^{t-d}_{-\infty}\right\}, \tag{4.19}$$

and satisfies $\sum_{n,k,m} I_{n,k,m} \, P^*_{n,k,m}(\mu^*_{\mathbf{I}}) = X_{\mathsf{con}}$.

Based on (4.16)-(4.19), Table 4.1 details the brute-force steps for a given $\mathbf{I}$. In the end, for a specified tolerance κ, these steps find $\underline{\mu}$ and $\bar{\mu}$ such that $\mu^*_{\mathbf{I}} \in [\underline{\mu}, \bar{\mu}]$ and $\bar{\mu} - \underline{\mu} < \kappa$. Using an approximation of $\mu^*_{\mathbf{I}}$ that lies in $[\underline{\mu}, \bar{\mu}]$, the corresponding utility is guaranteed to be no less than κX_{con} from the optimal (for the given $\mathbf{I}$). Therefore, by adjusting κ, one can achieve a performance arbitrarily close to the optimum. Since $|\mathcal{I}| = (KM + 1)^N$ values of $\mathbf{I}$ must be considered, the total complexity of the brute-force approach—in terms of the number of times (4.17) must be solved—can be shown to be

$$\left\lceil \log_2\left(\tfrac{\mu_{\mathsf{max}} - \mu_{\mathsf{min}}}{\kappa}\right) \right\rceil \times (KM + 1)^{N-1} NKM, \tag{4.20}$$

which grows exponentially with N.

4.5.2 Proposed Algorithm

We propose to attack the mixed-integer GSRA problem (4.12) using the well known *Lagrangian relaxation* approach [49]. In doing so, we relax the domain of the scheduling variables $I^t_{n,k,m}$ from the set $\{0, 1\}$ to the interval $[0, 1]$, allowing the application of low-complexity dual optimization techniques. Although the solution to the relaxed problem does not necessarily coincide with that of the original greedy

Table 4.1: Brute-force steps for a given $\mathbf{I}$

1. Initialize $\underline{\mu} = \mu_{\mathsf{min}}$ and $\bar{\mu} = \mu_{\mathsf{max}}$.
2. Set $\mu = \frac{\underline{\mu}+\bar{\mu}}{2}$.
3. For each (n, k, m),
   (a) Use (4.16)-(4.17) to obtain $P^*_{n,k,m}(\mu)$.
4. Calculate $X^*_{\mathsf{tot}}(\mathbf{I}, \mu) \triangleq \sum_{n,k,m} I_{n,k,m} P^*_{n,k,m}(\mu)$.
5. If $X^*_{\mathsf{tot}}(\mathbf{I}, \mu) > X_{\mathsf{con}}$, set $\underline{\mu} = \mu$, otherwise set $\bar{\mu} = \mu$.
6. If $\bar{\mu} - \underline{\mu} > \kappa$, go to step 2), else proceed to step 7).
7. If $X^*_{\mathsf{tot}}(\mathbf{I}, \bar{\mu}) \neq X^*_{\mathsf{tot}}(\mathbf{I}, \underline{\mu})$, set $\lambda = \frac{X^*_{\mathsf{tot}}(\mathbf{I},\underline{\mu}) - X_{\mathsf{con}}}{X^*_{\mathsf{tot}}(\mathbf{I},\underline{\mu}) - X^*_{\mathsf{tot}}(\mathbf{I},\bar{\mu})}$, otherwise set $\lambda = 0$.
8. Set $\hat{\mu}_{\mathbf{I}} = \bar{\mu}$. The best power allocation is given by $\hat{\boldsymbol{P}}(\mathbf{I}) = \lambda \mathbf{P}^*(\bar{\mu}) + (1-\lambda)\mathbf{P}^*(\underline{\mu})$, and $\hat{L}_{\mathbf{I}} = L^t_{\mathbf{I}}(\bar{\mu}, \hat{\boldsymbol{P}}(\mathbf{I}))$ gives the best Lagrangian value.

problem (4.12), we establish in the sequel that the corresponding performance loss is very small, and in some cases zero.

The relaxed version of the greedy problem (4.12) is

$$\begin{aligned} \text{rGSRA} \quad \triangleq \quad & \max_{\substack{\mathbf{I}^t \in \mathcal{I}_c \\ \mathbf{P}^t \in \mathcal{P}}} \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{m=1}^{M} I^t_{n,k,m} \, \mathrm{E}\left\{ U_{n,k,m}\big((1 - a_m e^{-b_m P^t_{n,k,m} \gamma_{n,k}}) r_m\big) \,\middle|\, \mathbb{F}^{t-d}_{-\infty} \right\} \\ & \text{s.t.} \quad \sum_{n,k,m} I^t_{n,k,m} \, P^t_{n,k,m} \leqslant X_{\mathsf{con}}, \end{aligned} \tag{4.21}$$

where $\mathcal{I}_c \triangleq \big\{ \mathbf{I} \in [0,1]^{NKM} : \sum_{k,m} I_{n,k,m} \leqslant 1 \ \forall n \big\}$. Although (4.21) is a non-convex optimization problem due to non-convex constraints, it can be converted into a convex optimization problem by using the new set of variables $(\mathbf{I}^t, \mathbf{x}^t)$, where $x^t_{n,k,m} \triangleq I^t_{n,k,m} \, P^t_{n,k,m}$. In this case, we have

$$\text{rGSRA} = \min_{\substack{\mathbf{x}^t \succeq 0 \\ \mathbf{I}^t \in \mathcal{I}_c}} \sum_{n,k,m} I^t_{n,k,m} \, B^t_{n,k,m}(I^t_{n,k,m}, x^t_{n,k,m}) \quad \text{s.t.} \sum_{n,k,m} x^t_{n,k,m} \leqslant X_{\mathsf{con}}, \tag{4.22}$$

where $\mathbf{x}^t \in \mathbb{R}^{NKM}$ denotes the collection of all time-t variables $\{x^t_{n,k,m}\}$, $\mathbf{x}^t \succeq 0$ denotes element-wise non-negativity, and $B^t_{n,k,m}(\cdot,\cdot)$ is defined as

$$B^t_{n,k,m}(y_1,y_2) \;\triangleq\; \begin{cases} -\,\mathrm{E}\left\{U_{n,k,m}\big((1-a_m e^{-b_m\gamma^t_{n,k}\,y_2/y_1})r_m\big)\Big|\mathbb{F}^{t-d}_{-\infty}\right\} & \text{if } y_1 \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{4.23}$$

The modified problem (4.22) is a convex optimization problem and can be solved using a dual optimization approach with zero duality gap. In particular, the dual problem can be written as

$$\begin{aligned} &\max_{\mu\geqslant 0}\min_{\substack{\mathbf{x}^t\succeq 0\\ \mathbf{I}^t\in\mathcal{I}_c}} L(\mu,\mathbf{I}^t,\mathbf{x}^t) = \max_{\mu\geqslant 0}\min_{\mathbf{I}^t\in\mathcal{I}_c} L(\mu,\mathbf{I}^t,\mathbf{x}^{t,*}(\mu,\mathbf{I}^t)) \\ &\quad= \max_{\mu\geqslant 0} L(\mu,\mathbf{I}^{t,*}(\mu),\mathbf{x}^{t,*}(\mu,\mathbf{I}^{t,*}(\mu))) = L(\mu^*,\mathbf{I}^{t,*}(\mu^*),\mathbf{x}^{t,*}(\mu^*,\mathbf{I}^{t,*}(\mu^*))), \end{aligned} \tag{4.24}$$

where

$$L(\mu,\mathbf{I}^t,\mathbf{x}^t) \triangleq \sum_{n,k,m} I^t_{n,k,m}\, B^t_{n,k,m}(I^t_{n,k,m},x^t_{n,k,m}) + \Big(\sum_{n,k,m} x^t_{n,k,m} - X_{\mathsf{con}}\Big)\mu, \tag{4.25}$$

where $\mathbf{x}^*(\mu,\mathbf{I})$ is the optimal $\mathbf{x}$ for a given $(\mu,\mathbf{I})$, where $\mathbf{I}^*(\mu)$ denotes the optimal $\mathbf{I}\in\mathcal{I}_c$ for a given μ, and where μ^* denotes the optimal $\mu\geqslant 0$.

A detailed solution to this problem is given in Chapter 3, and so we describe only the main points here. For given values of μ and $\mathbf{I}^t$, we have $x^{t,*}_{n,k,m}(\mu,\mathbf{I}^t) = I^t_{n,k,m}\,P^{t,*}_{n,k,m}(\mu)$, where

$$P^{t,*}_{n,k,m}(\mu) \;=\; \begin{cases} \tilde{P}^t_{n,k,m}(\mu) & \text{if } 0\leqslant\mu\leqslant a_m b_m r_m U'_{n,k,m}\big((1-a_m)r_m\big)\,\mathrm{E}\left\{\gamma^t_{n,k}\big|\mathbb{F}^{t-d}_{-\infty}\right\} \\ 0 & \text{otherwise,} \end{cases} \tag{4.26}$$

and where $\tilde{P}^t_{n,k,m}(\mu)$ is defined as the (unique) solution to

$$\mu \;=\; a_m b_m r_m\,\mathrm{E}\left\{U'_{n,k,m}\big((1-a_m e^{-b_m\tilde{P}^t_{n,k,m}(\mu)\gamma^t_{n,k}})r_m\big)\gamma^t_{n,k}e^{-b_m\tilde{P}^t_{n,k,m}(\mu)\gamma^t_{n,k}}\big|\mathbb{F}^{t-d}_{-\infty}\right\}. \tag{4.27}$$

To give equations that govern $\mathbf{I}^{t,*}(\mu)$ for a given μ, we first define

$$\begin{aligned} V^t_{n,k,m}(\mu, P^{t,*}_{n,k,m}(\mu)) \;\triangleq\; & -\mathrm{E}\left\{U_{n,k,m}\big((1-a_m e^{-b_m P^{t,*}_{n,k,m}(\mu)\gamma_{n,k}})r_m\big)\big|\mathbb{F}^{t-d}_{-\infty}\right\} \\ & + \mu P^*_{n,k,m}(\mu) \end{aligned} \tag{4.28}$$

and

$$S^t_n(\mu) \triangleq \Big\{(k,m) = \underset{(k',m')}{\operatorname{argmin}} V^t_{n,k',m'}(\mu, P^{t,*}_{n,k',m'}(\mu)) : V^t_{n,k,m}(\mu, P^{t,*}_{n,k,m}(\mu)) \leqslant 0\Big\}. \tag{4.29}$$

If $S^t_n(\mu)$ is a null or a singleton set, then the optimal schedule on subchannel n is given by

$$I^{t,*}_{n,k,m}(\mu) = \begin{cases} 1 & (k,m) \in S^t_n(\mu) \\ 0 & \text{otherwise.} \end{cases} \tag{4.30}$$

However, if $S^t_n(\mu)$ has cardinality greater than one, then multiple (k,m) combinations can be scheduled simultaneously while achieving the optimal value of the Lagrangian. In particular, if $S^t_n(\mu) = \{(k_1(n), m_1(n)), \ldots, (k_{|S^t_n(\mu)|}(n), m_{|S^t_n(\mu)|}(n))\}$, then

$$I^{t,*}_{n,k,m}(\mu) = \begin{cases} I_{n,k_i(n),m_i(n)} & \text{if } (k,m) = (k_i(n), m_i(n)) \text{ for some } i \in \{1, \ldots, |S^t_n(\mu)|\} \\ 0 & \text{otherwise,} \end{cases} \tag{4.31}$$

where the vector $\left[I_{n,k_1(n),m_1(n)}, \ldots, I_{n,k_{|S^t_n(\mu)|}(n),m_{|S^t_n(\mu)|}(n)}\right]$ lies anywhere in the unit-$(|S^t_n(\mu)|-1)$ simplex, i.e., it lies within the region $[0,1]^{|S^t_n(\mu)|}$ and satisfies the equation $\sum_{i=1}^{|S^t_n(\mu)|} I_{n,k_i(n),m_i(n)} = 1$. Finally, the optimal Lagrange multiplier μ (i.e., μ^*) is such that $\mu^* \in [\mu_{\mathsf{min}}, \mu_{\mathsf{max}}] \subset (0, \infty)$ and

$$\sum_{n,k,m} I^{t,*}_{n,k,m}(\mu^*)\, P^{t,*}_{n,k,m}(\mu^*) = X_{\mathsf{con}}, \tag{4.32}$$

where μ_{min} and μ_{max} were given in (4.18) and (4.19), respectively.

For several fixed values of μ, the proposed algorithm minimizes the relaxed Lagrangian (4.25) over $(\mathbf{I}^t, \mathbf{x}^t)$ (or, equivalently, over $(\mathbf{I}^t, \mathbf{P}^t)$) to obtain candidate solutions for the original greedy problem (4.12). If, for a given μ, $|S_n^t(\mu)| \leqslant 1$ for all n (i.e., the candidate employs at most one user/MCS per subchannel), then the candidate solution is admissible for the non-relaxed problem, and thus retained by the proposed algorithm. If, on the other hand, $|S_n^t(\mu)| > 1$ for some n (i.e., the candidate employs more than one user/MCS on some subchannels), then the proposed algorithm transforms the candidate into an admissible solution as follows:

$$I_{n,k,m}^{t,\mathsf{pro}}(\mu) = \begin{cases} 1 & (k,m) = \operatorname{argmin}_{(k',m') \in S_n^t(\mu)} P_{n,k',m'}^{t,*}(\mu) \\ 0 & \text{otherwise.} \end{cases} \tag{4.33}$$

The following lemma then states an important property of these fixed-μ admissible solutions.

Lemma 8. *For any given value of μ, let the power allocation $\mathbf{P}^{t,*}(\mu)$ be given by (4.26), let the user-MCS allocation $\mathbf{I}^{t,\mathsf{pro}}(\mu)$ be given by (4.33), and let the total power allocation be defined as $X_{\mathsf{tot}}^{t,\mathsf{pro}}(\mu) \triangleq \sum_{n,k,m} I_{n,k,m}^{t,\mathsf{pro}}(\mu)\, P_{n,k,m}^{t,*}(\mu)$. Then, $X_{\mathsf{tot}}^{t,\mathsf{pro}}(\mu)$ is monotonically decreasing in μ.*

Lemma 8 (see Chapter 3 for a proof) implies that the optimal value of the Lagrange multiplier μ (i.e., μ^*) is the one that achieves the power constraint $X_{\mathsf{tot}}^{t,\mathsf{pro}}(\mu) = X_{\mathsf{con}}$. To find this μ^*, the proposed algorithm performs a bisection search over $\mu \in [\mu_{\mathsf{min}}, \mu_{\mathsf{max}}]$ that refines the search interval $[\underline{\mu}, \bar{\mu}]$ until $\bar{\mu} - \underline{\mu} < \kappa$, where κ is a user-defined tolerance. Then, between the two schedules $\mathbf{I} \in \{\mathbf{I}^{t,\mathsf{pro}}(\underline{\mu}), \mathbf{I}^{t,\mathsf{pro}}(\bar{\mu})\}$, it chooses the one that maximizes utility, reminiscent of the brute-force algorithm. Table 4.2 summarizes the proposed algorithm.

The complexity of the proposed algorithm—in terms of number of times (4.27) is solved—is

$$\left\lceil \log_2\left(\tfrac{\mu_{\mathsf{max}} - \mu_{\mathsf{min}}}{\kappa}\right) \right\rceil \times N(KM + 2), \tag{4.34}$$

Table 4.2: Proposed greedy algorithm

1. Initialize $\underline{\mu} = \mu_{\mathsf{min}}$ and $\bar{\mu} = \mu_{\mathsf{max}}$.
2. Set $\mu = \frac{\underline{\mu}+\bar{\mu}}{2}$.
3. For each subchannel $n = 1, \ldots, N$:
   (a) For each (k, m),
      i. Use (4.26)-(4.27) to calculate $P^{t,*}_{n,k,m}(\mu)$.
      ii. Use $P^{t,*}_{n,k,m}(\mu)$ to calculate $V^t_{n,k,m}(\mu, P^{t,*}_{n,k,m}(\mu))$ via (4.28).
   (b) Calculate $S^t_n(\mu)$ using (4.29).
4. Find $\mathbf{I}^{t,\mathsf{pro}}(\mu)$ using (4.33).
5. Calculate $X^{t,\mathsf{pro}}_{\mathsf{tot}}(\mu) = \sum_{n,k,m} I^{t,\mathsf{pro}}_{n,k,m}(\mu)\, P^{t,*}_{n,k,m}(\mu)$.
6. If $X^{t,\mathsf{pro}}_{\mathsf{tot}}(\mu) > X_{\mathsf{con}}$, set $\underline{\mu} = \mu$, otherwise set $\bar{\mu} = \mu$.
7. If $\bar{\mu} - \underline{\mu} > \kappa$, go to step 2), else proceed to step 8).
8. Now we have $\mu^* \in [\underline{\mu}, \bar{\mu}]$ and $\bar{\mu} - \underline{\mu} < \kappa$. For both $\mathbf{I} = \mathbf{I}^{t,\mathsf{pro}}(\underline{\mu})$ and $\mathbf{I} = \mathbf{I}^{t,\mathsf{pro}}(\bar{\mu})$ (since they may differ), calculate $\hat{\boldsymbol{P}}(\mathbf{I})$ and $\hat{L}_{\mathbf{I}}$ as described for the brute force algorithm.
9. Choose $\hat{\boldsymbol{I}}^t = \mathrm{argmin}_{\mathbf{I}\in\{\mathbf{I}^{t,\mathsf{pro}}(\underline{\mu}),\, \mathbf{I}^{t,\mathsf{pro}}(\bar{\mu})\}} \hat{L}_{\mathbf{I}}$ as the user-MCS allocation and $\hat{\boldsymbol{P}}^t = \hat{\boldsymbol{P}}(\hat{\boldsymbol{I}}^t)$ as the associated power allocation.

which is significantly less than the brute-force complexity in (4.20). Although the proposed algorithm is sub-optimal, the difference between the optimal GSRA utility U^*_{GSRA} and that attained by the proposed algorithm $\hat{U}_{\mathsf{GSRA}}(\underline{\mu}, \bar{\mu})$, as $\underline{\mu} \to \bar{\mu}$, can be bounded as follows (see Chapter 3 for details):

$$U^*_{\mathsf{GSRA}} - \lim_{\underline{\mu}\to\bar{\mu}} \hat{U}_{\mathsf{GSRA}}(\underline{\mu}, \bar{\mu}) \;\leqslant\; (\mu^* - \mu_{\mathsf{min}})\big(X_{\mathsf{con}} - X^{t,\mathsf{pro}}_{\mathsf{tot}}(\mu^*)\big) \tag{4.35}$$

$$\leqslant \begin{cases} 0 & \text{if } |S_n(\mu^*)| \leqslant 1\ \forall n \\ (\mu_{\mathsf{max}} - \mu_{\mathsf{min}})X_{\mathsf{con}} & \text{otherwise.} \end{cases} \tag{4.36}$$

In Section 4.7, we evaluate (4.35) by simulation, and show that the performance loss is negligible.

4.6 Updating the Posterior Distributions from ACK/NAK Feedback

In this section, we propose a recursive procedure to compute the posterior pdfs $p(\gamma_{n,k}^t \,|\, \mathbb{F}_{-\infty}^{t-d})$ required by the proposed greedy algorithm in Table 4.2 when the channel is first-order[16] Markov.

Let the time-t user-k channel be described by the discrete-time channel impulse response $\mathbf{h}_k^t \triangleq [h_{1,k}^t, \ldots, h_{L,k}^t]^\intercal \in \mathbb{C}^L$, where $(\cdot)^\intercal$ denotes transpose. The corresponding frequency-domain subchannel gains $\mathbf{H}_k^t \triangleq [H_{1,k}^t, \ldots, H_{N,k}^t]^\intercal \in \mathbb{C}^N$ are then given by

$$\mathbf{H}_k^t \;=\; \mathbf{G}\mathbf{h}_k^t, \tag{4.37}$$

where the OFDMA modulation matrix $\mathbf{G} \in \mathbb{C}^{N\times L}$ contains the first L columns of the N-DFT matrix. Assuming additive white Gaussian noise with unit variance, the SSG of subchannel n for user k is given by $\gamma_{n,k}^t = |H_{n,k}^t|^2$, and so we can write

$$p(\gamma_{n,k}^t \,|\, \mathbb{F}_{-\infty}^{t-d}) = \int_{\mathbf{h}_k^t} p(\gamma_{n,k}^t \,|\, \mathbf{h}_k^t) p(\mathbf{h}_k^t \,|\, \mathbb{F}_{-\infty}^{t-d}) \tag{4.38}$$

with $p(\gamma_{n,k}^t \,|\, \mathbf{h}_k^t) = \delta(\gamma_{n,k}^t - |\mathbf{e}_n^\intercal \mathbf{G}\mathbf{h}_k^t|^2)$, where $\delta(\cdot)$ is the Dirac delta and $\mathbf{e}_n$ is the n^{th} column of the identity matrix. Using the channel's Markov property and Bayes rule, we find that

$$p(\mathbf{h}_k^t \,|\, \mathbb{F}_{-\infty}^{t-d}) \;=\; \int_{\mathbf{h}_k^{t-d}} p(\mathbf{h}_k^t \,|\, \mathbf{h}_k^{t-d})\, p(\mathbf{h}_k^{t-d} \,|\, \mathbb{F}_{-\infty}^{t-d}) \tag{4.39}$$

$$p(\mathbf{h}_k^{t-d} \,|\, \mathbb{F}_{-\infty}^{t-d}) \;=\; \frac{p(\mathbf{f}_k^{t-d} \,|\, \mathbf{h}_k^{t-d}, \mathbb{F}_{-\infty}^{t-d} \setminus \mathbf{f}_k^{t-d})\, p(\mathbf{h}_k^{t-d} \,|\, \mathbb{F}_{-\infty}^{t-d} \setminus \mathbf{f}_k^{t-d})}{\int_{\bar{\mathbf{h}}_k^{t-d}} p(\mathbf{f}_k^{t-d} \,|\, \bar{\mathbf{h}}_k^{t-d}, \mathbb{F}_{-\infty}^{t-d} \setminus \mathbf{f}_k^{t-d})\, p(\bar{\mathbf{h}}_k^{t-d} \,|\, \mathbb{F}_{-\infty}^{t-d} \setminus \mathbf{f}_k^{t-d})}, \tag{4.40}$$

where $\setminus$ denotes the set-difference operator. Using the fact that $p(\mathbf{f}_k^{t-d} \,|\, \mathbf{h}_k^{t-d}, \mathbb{F}_{-\infty}^{t-d} \setminus \mathbf{f}_k^{t-d}) = p(\mathbf{f}_k^{t-d} \,|\, \mathbf{h}_k^{t-d}, \mathbf{I}^{t-d}, \mathbf{p}^{t-d})$, along with the fact that $(\mathbf{I}^{t-d}, \mathbf{p}^{t-d})$ is a deterministic

[16] The extension to higher-order Markov channels is straightforward.

function of $\mathbb{F}_{-\infty}^{t-2d}$ (and therefore of $\mathbb{F}_{-\infty}^{t-d-1}$), we then have from (4.40) that

$$p(\mathbf{h}_k^{t-d} \,|\, \mathbb{F}_{-\infty}^{t-d}) \;=\; \frac{p(\mathbf{f}_k^{t-d} \,|\, \mathbf{h}_k^{t-d}, \mathbf{I}^{t-d}, \mathbf{p}^{t-d})\, p(\mathbf{h}_k^{t-d} \,|\, \mathbb{F}_{-\infty}^{t-d-1})}{\int_{\bar{\mathbf{h}}_k^{t-d}} p(\mathbf{f}_k^{t-d} \,|\, \bar{\mathbf{h}}_k^{t-d}, \mathbf{I}^{t-d}, \mathbf{p}^{t-d})\, p(\bar{\mathbf{h}}_k^{t-d} \,|\, \mathbb{F}_{-\infty}^{t-d-1})}. \tag{4.41}$$

Using the Markov property again, we get

$$p(\mathbf{h}_k^{t-d} \,|\, \mathbb{F}_{-\infty}^{t-d-1}) \;=\; \int_{\mathbf{h}_k^{t-d-1}} p(\mathbf{h}_k^{t-d} \,|\, \mathbf{h}_k^{t-d-1})\, p(\mathbf{h}_k^{t-d-1} \,|\, \mathbb{F}_{-\infty}^{t-d-1}). \tag{4.42}$$

Recall that $f_{n,k}^t$, the feedback received about user k on channel n at time t, takes values from the set $\{0, 1, \emptyset\}$, where 0 denotes a NAK, 1 denotes an ACK, and $\emptyset$ denotes no feedback. Here, $f_{n,k}^t$ is set to $\emptyset$ if user k was not scheduled on subchannel n at time t. Assuming that, conditioned on $\mathbf{h}_k^t$, the feedbacks generated by user k are independent across subchannels, we have

$$p(\mathbf{f}_k^t | \mathbf{h}_k^t, \mathbf{I}^t, \mathbf{p}^t) = \prod_{n=1}^{N} p(f_{n,k}^t | \mathbf{h}_k^t, \mathbf{I}^t, \mathbf{p}^t), \tag{4.43}$$

$$p(f_{n,k}^t = f \,|\, \mathbf{h}_k^t, \mathbf{I}^t, \mathbf{p}^t) = \begin{cases} \sum_m I_{n,k,m}^t a_m e^{-b_m p_{n,k,m}^t \gamma_{n,k}^t} & \text{if } f = 0 \\ \sum_m I_{n,k,m}^t \left(1 - a_m e^{-b_m p_{n,k,m}^t \gamma_{n,k}^t}\right) & \text{if } f = 1 \\ 1 - \sum_m I_{n,k,m}^t & \text{if } f = \emptyset, \end{cases} \tag{4.44}$$

where $\gamma_{n,k}^t = |H_{n,k}^t|^2$ can be determined from $\mathbf{h}_k^t$ via (4.37). Together, (4.38)-(4.44) suggest a method of *recursively* updating the channel distributions, using the new feedback obtained at each time t, which is given in Table 4.3.

We now propose the use of particle filtering [65] to circumvent the evaluation of multidimensional integrals in the recursion of Table 4.3. Particle filtering is a well-known technique that approximates the pdf of a random variable using a suitably chosen probability mass function (pmf). In the sequel, for simplicity of illustrations, we assume a Gauss-Markov model of the form

$$h_{l,k}^{t+1} \;=\; (1 - \alpha) h_{l,k}^t + \alpha w_{l,k}^t, \tag{4.45}$$

Table 4.3: Recursive update of channel posteriors

At time t, for each user k, the pdf $p(\mathbf{h}_k^{t-d-1} \,|\, \mathbb{F}_{-\infty}^{t-d-1})$ is available from the previous time-instant. The user-k recursion is then

1. Observe new feedbacks $\mathbf{f}_k^{t-d} \in \{0, 1, \emptyset\}^N$.
2. Compute $p(\mathbf{h}_k^{t-d} \,|\, \mathbb{F}_{-\infty}^{t-d-1})$ using (4.42).
3. Compute $p(\mathbf{f}_k^{t-d} \,|\, \mathbf{h}_k^{t-d}, \mathbf{I}^{t-d}, \mathbf{P}^{t-d})$ using the error-rate rule (4.43)-(4.44).
4. Using the distributions obtained in steps 2) and 3), compute $p(\mathbf{h}_k^{t-d} \,|\, \mathbb{F}_{-\infty}^{t-d})$ via Bayes-rule step in (4.41).
5. Compute $p(\mathbf{h}_k^{t} \,|\, \mathbb{F}_{-\infty}^{t-d})$ using the Markov-prediction step (4.39).
6. For each n, compute $p(\gamma_{n,k}^{t} \,|\, \mathbb{F}_{-\infty}^{t-d})$ via (4.38).

where $w_{l,k}^t$ is unit-variance circular Gaussian and $\alpha \in (0, 1]$ is a known constant that determines the fading rate. Here, $w_{l,k}^t$ is assumed to be i.i.d. for all t, l, k. At each time-step t, for $k \in \{1, \ldots K\}$, we use S particles in the approximations

$$\begin{aligned} p(\mathbf{h}_k^t \,|\, \mathbb{F}_{-\infty}^{t-d}) &\approx \sum_{i=1}^{S} \nu_k^{t\,|\,t-d}[i]\, \delta(\mathbf{h}_k^t - \mathbf{h}_k^t[i]), \text{ and} \\ p(\mathbf{h}_k^{t-d} \,|\, \mathbb{F}_{-\infty}^{t-d}) &\approx \sum_{i=1}^{S} \nu_k^{t-d\,|\,t-d}[i]\, \delta(\mathbf{h}_k^{t-d} - \mathbf{h}_k^{t-d}[i]), \end{aligned} \tag{4.46}$$

where $\mathbf{h}_k^t[i] = \big[h_{1,k}^t[i], \ldots, h_{L,k}^t[i]\big]^\intercal \in \mathbb{C}^L$ denotes the i^{th} (vector) particle, for $i \in \{1, \ldots, S\}$, and $\nu_k^{t_1|t_2}[i] \in \mathbb{R}^+$ is the probability mass assigned to the particle $\mathbf{h}_k^{t_1}[i]$ based on the observations received up to time t_2. The steps to recursively compute these particles and their corresponding weights are detailed in Table 4.4.

Using the approximation in (4.46), we note that the expectation of any function of subchannel-gain, $\mathbf{h}_k^t$, can be found using

$$\mathrm{E}\{A(\mathbf{h}_k^t)\} \approx \sum_i \nu_k^{t\,|\,t-d}[i] A\big(\mathbf{h}_k^t[i]\big), \tag{4.47}$$

where $A(\cdot)$ is an arbitrary function. Recalling that the SSG $\gamma_{n,k}^t$ is a deterministic function of the subchannel-gain, $\mathbf{h}_k^t$, any function of $\gamma_{n,k}^t$ is also a function of $\mathbf{h}_k^t$.

4.7 Numerical Results

In this section, we numerically evaluate the performance of the proposed greedy scheduling and resource allocation from Section 4.5 with the posterior update from Section 4.6. For this, we consider an OFDMA system with independent first-order Gauss-Markov channels (4.45). We assumed, if not otherwise stated, $K = 8$ available users, $N = 32$ OFDMA subchannels, channel fading parameter $\alpha = 10^{-3}$ and impulse response length $L = 2$. We used the modulation matrix $\mathbf{G} = \sqrt{\beta}\mathbf{F} \in \mathbb{C}^{N\times L}$ (recall (4.37)), where $\mathbf{F}$ contains the first $L(\leqslant N)$ columns of the unitary N-DFT matrix and $\beta = \frac{N}{L}\frac{2-\alpha}{\alpha}$ ensures that the variance of $H^t_{n,k}$ is unity for all (n,k). Thus, the mean of the SSG $\gamma^t_{n,k}$ was also unity for all (n,k). Since the subchannel-averaged total transmit power equals $\frac{1}{N}\sum_{n,k,m} I^t_{n,k,m}P^t_{n,k,m} = \frac{1}{N}\sum_{n,k,m} X^t_{n,k,m} = X_{\mathsf{con}}/N$, it is readily seen that the average per-subchannel signal-to-noise ratio is $\mathsf{SNR} \triangleq \mathrm{E}\{\frac{1}{N}\sum_{n,k,m} X^t_{n,k,m}\gamma^t_{n,k}\} = \frac{1}{N}\sum_{n,k,m} X^t_{n,k,m}\,\mathrm{E}\{\gamma^t_{n,k}\} = X_{\mathsf{con}}/N$. For the plots, we averaged 500 realizations, each with 100 time-slots. Of these 100 time-slots, the first 50 were ignored to avoid transient effects.

For illustrative purposes, we assumed uncoded 2^{m+1}-QAM signaling with MCS index $m \in \{1,\ldots,15\}$. In this case, we have $r_m = m+1$ bits per symbol, one symbol per "codeword," and one codeword per packet. In the packet error-rate model $\epsilon = a_m e^{-b_m P\gamma}$, we assumed $a_m = 1$ and $b_m = 1.5/(2^{m+1}-1)$ because the symbol error-rate of a 2^{m+1}-QAM system is well approximated by $\exp(-1.5P\gamma/(2^{m+1}-1))$ in the high-$(P\gamma)$ regime [51] and is ≈ 1 when $P\gamma = 0$. Throughout, we used the identity utility (i.e., $U_{n,k,m}(x) = x$ for all n,k,m) so that the objective was maximization of sum goodput, and we assumed a feedback delay of $d = 1$.

The performance of the proposed greedy algorithm was compared to three reference schemes: fixed-power random user scheduling (FP-RUS), the "causal global genie" (CGG), and the "non-causal global genie" (NCGG). The FP-RUS scheme schedules users uniformly at random, allocates power uniformly across subchannels, and selects the MCS to maximize expected goodput. The FP-RUS, which makes no use of feedback, should perform no better than any feedback-based scheme. The CGG (recall Section 4.4.1) performs optimal scheduling and resource allocation under perfect knowledge of all SSGs at the previous time-instant (since $d = 1$), i.e., given $\{\gamma_{n,k}^{t-1} \ \forall n, k\}$ at time t. From Lemma 7, we know that the CGG upper-bounds the POMDP. The NCGG is similar to the CGG, but assumes perfect knowledge of all SSGs at *all* times, i.e., given $\{\gamma_{n,k}^{\tau} \ \forall n, k, \tau\}$ at time t. Thus, it provides an upper bound on the CGG that is invariant to fading rate α. The NCGG has a greedy implementation, like the CGG, but without the conditional expectation in (4.5).

Figure 4.1 shows a typical realization of instantaneous sum-goodput versus time t, when $\alpha = 10^{-3}$. There, one can see a large gap between the FP-RUS and the CGG, and a much smaller gap between the CGG and the NCGG. The proposed scheme starts without CSI, and initially performs no better than the FP-RUS. From ACK/NAK feedbacks, however, it quickly learns the CSI well enough to perform scheduling and resource allocation at a level that yields sum-goodput much closer to the CGG than to the FP-RUS.

Figure 4.2 plots average sum-goodput versus the number of particles S used to update the posterior distributions in the proposed greedy scheme (recall Section 4.6). There we see that the performance of the proposed scheme increases with S, but shows little improvement for $S > 30$. Thus, $S = 30$ particles were used to construct the

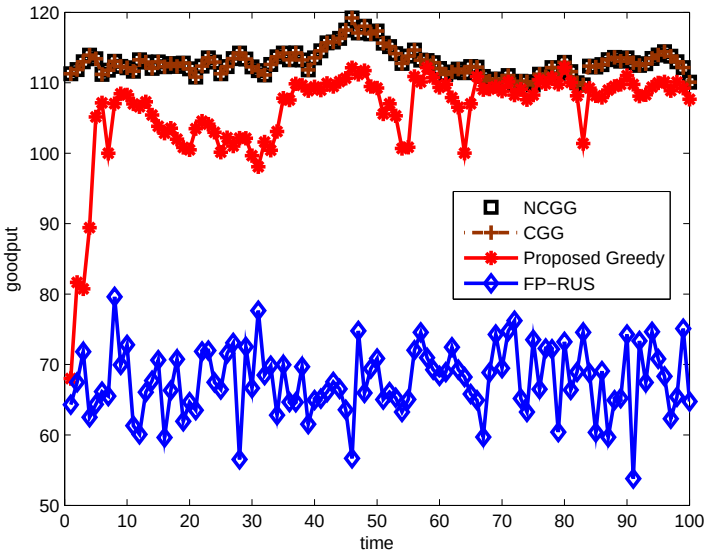

Figure 4.1: Typical instantaneous sum-goodput versus time t. Here, $N = 32$, $K = 8$, $\mathsf{SNR} = 10\text{dB}$, $\alpha = 10^{-3}$, and $S = 30$.

other plots. Remarkably, with only $S = 5$ particles, the proposed algorithm captures a significant portion of the maximum possible goodput gain over the FP-RUS.

Figure 4.3 plots average sum-goodput versus the fading rate α. There we see that, at low fading rates (i.e., small α), the proposed greedy scheme achieves an average sum-goodput that is much higher than the FP-RUS and, in fact, not far from the CGG upper bound. For instance, at $\alpha = 10^{-4}$, the sum-goodput attained by the proposed scheme is 92% of the upper bound and 170% of that attained by the FP-RUS. As the fading rate α increases, we see that the sum-goodput attained by the proposed scheme decreases, and eventually converges to that of the FP-RUS. This behavior is due to the fact that, as α increases, it becomes more difficult to predict the SSGs using delayed ACK/NAK feedback, thereby compromising the scheduling-and-resource-allocation decisions that are made based on the predicted SSGs. In fact,

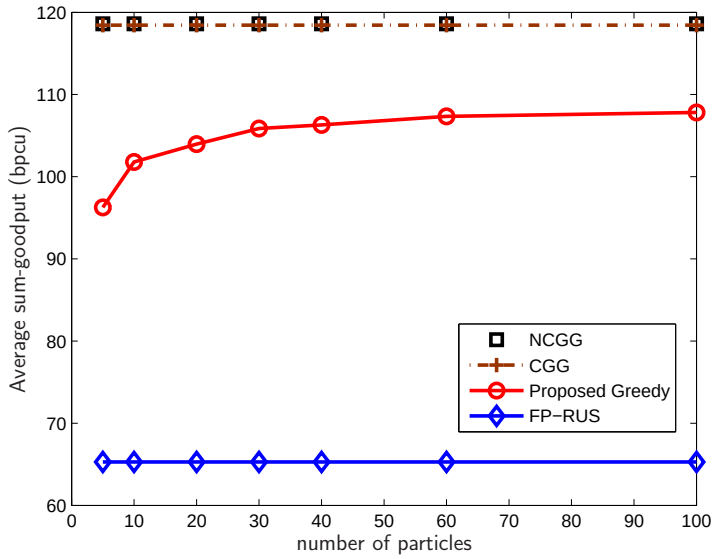

Figure 4.2: Average sum-goodput versus the number of particles used to update the channel posteriors. Here, $N = 32$, $K = 8$, $\mathsf{SNR} = 10\text{dB}$, and $\alpha = 10^{-3}$.

one can even observe a gap between the CGG and NCGG for large α because, even with delayed perfect-SSG feedback, the current SSGs are difficult to predict.

Figure 4.3 reveals a gap between the proposed scheme and the CGG bound that persists as $\alpha \to 0$. This non-vanishing gap can be attributed—at least in part—to *greedy* scheduling under ACK/NAK feedback. Intuitively, we have the following explanation. Because the inferred SSG-distributions of not-recently-scheduled users quickly revert to their a priori form, the proposed greedy algorithm will continue to schedule users as long as their SSGs remain better than the a priori value. There may exist, however, not-recently-scheduled users with far better SSGs who remain invisible to the proposed scheme, only because they have not recently been scheduled.

Figures 4.4 and 4.5 plot average sum-goodput versus the number of subchannels (i.e., total bandwidth) N. In Fig. 4.4, the total BS power X_{con} is scaled with N such

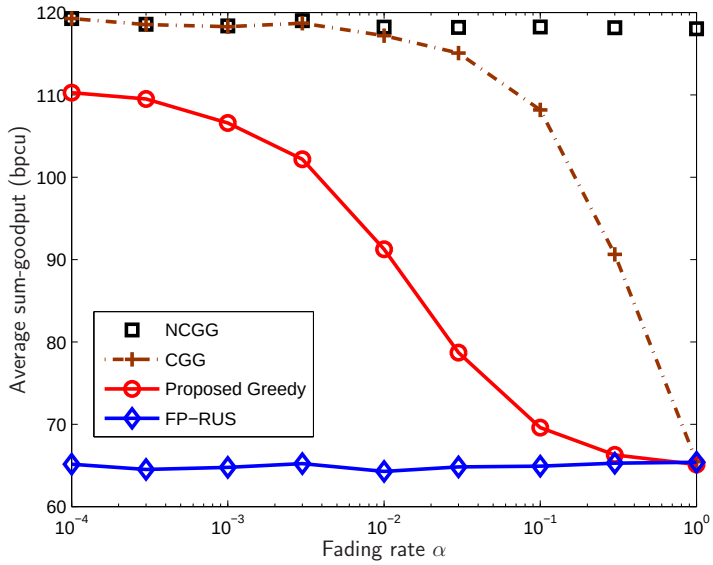

Figure 4.3: Average sum-goodput versus fading rate α. Here, $N = 32$, $K = 8$, $\mathsf{SNR} = 10\text{dB}$, and $S = 30$.

that the per-subchannel SNR remains fixed at 10dB, whereas, in Fig. 4.5, the total BS power X_{con} remains invariant to the bandwidth N, and is set such that per-subchannel $\mathsf{SNR} = 10\text{dB}$ for $N = 32$. In both cases, the average sum-goodput increases with bandwidth N, as expected, since the availability of more subchannels increases not only scheduling flexibility, but also the possibility of stronger subchannels, which can be exploited by the BS. In Fig. 4.4, where the per-subchannel SNR is fixed, the sum-goodput increases linearly with bandwidth N, as expected. In all cases, the proposed greedy scheme achieves more than 155% of the sum-goodput achieved by the FP-RUS.

Figure 4.6 plots average sum-goodput versus the number of available users K. It shows that, as K increases, the average sum-goodputs achieved by the NCGG, CGG, and the proposed greedy schemes increase, whereas that achieved by the FP-RUS

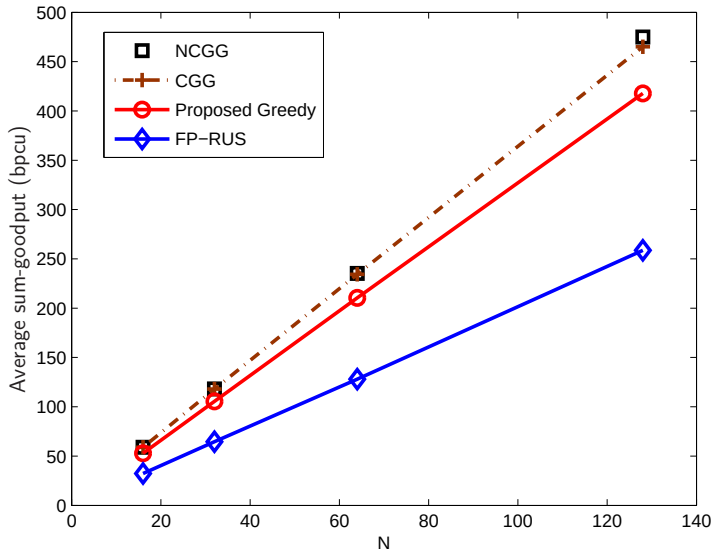

Figure 4.4: Average sum-goodput versus number of subchannels N. Here, $K = 8$, $\mathsf{SNR} = 10\text{dB}$, $\alpha = 10^{-3}$, and $S = 30$.

remains constant. This behavior results because, with the former schemes, the availability of more users can be exploited to schedule users with stronger subchannels, whereas with the FP-RUS scheme, this advantage is lost due to the complete lack of information about the users' instantaneous channel conditions. Figure 4.6 also suggests that, as K increases, the sum-goodput of the proposed greedy scheme saturates. This can be attributed to the fact that the proposed greedy algorithm can only track the channels of recently scheduled users, and thus cannot benefit directly from the growing pool of not-recently-scheduled users.

In Figure 4.7, the top subplot shows average sum-goodput versus SNR, while the bottom subplot shows the average value of the bound (4.35) on the optimality gap of our proposed approach to the GSRA problem, also versus SNR. The top plot shows that, as the SNR increases, the proposed greedy scheme continues to perform much

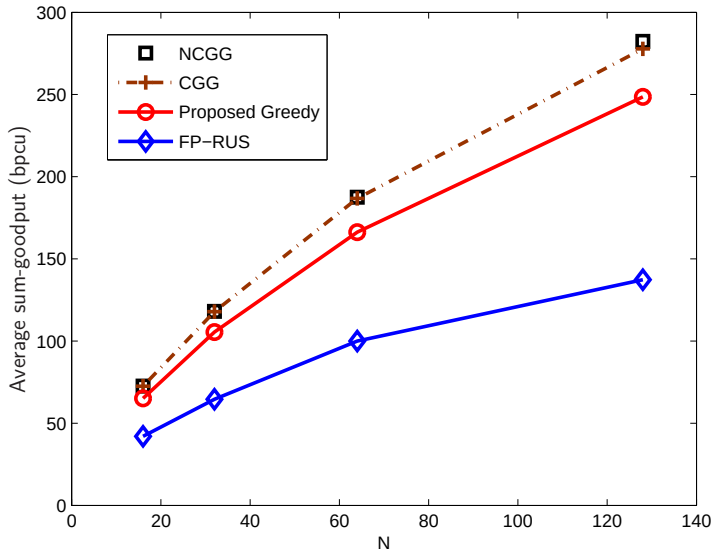

Figure 4.5: Average sum-goodput versus number of subchannels N. Here, $K = 8$, X_{con} does not scale with N and it is chosen such that $\mathsf{SNR} = 10\text{dB}$ for $N = 32$, $\alpha = 10^{-3}$, and $S = 30$.

closer to the NCGG/CGG bounds than it does to the FP-RUS scheme. The bottom plot establishes that the sum-goodput loss due to the sub-optimality in the algorithm used to attack the GSRA problem is negligible, e.g., at most 0.0025% over all SNR.

4.8 Summary

In this chapter, we considered the problem of joint scheduling and resource allocation in the OFDMA downlink under ACK/NAK feedback, with the goal of maximizing an expected long-term goodput-based utility subject to an instantaneous sum-power constraint. First, we established that the optimal solution to the problem is a partially observable Markov decision process (POMDP), which is impractical to implement. Consequently, we proposed a greedy approach to joint scheduling and

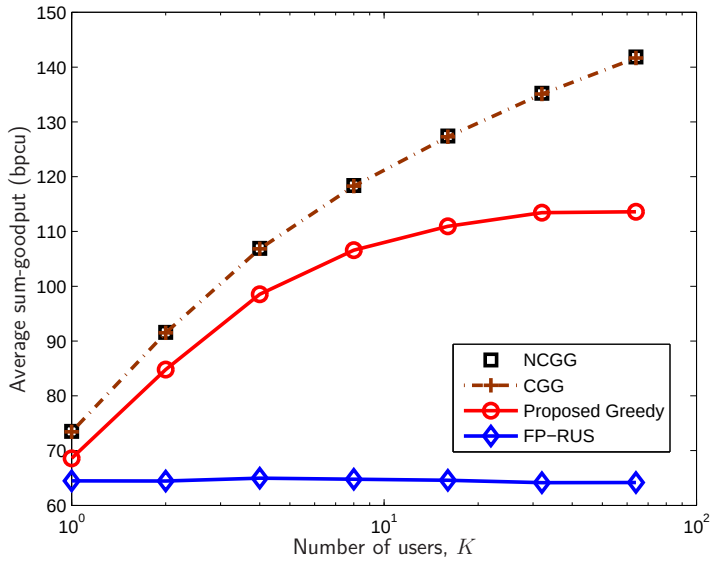

Figure 4.6: Average sum-goodput versus number of users. In this plot, $N = 32$, $\mathsf{SNR} = 10\text{dB}$, $\alpha = 10^{-3}$, and $S = 30$.

resource allocation based on the posterior distributions of the squared subchannel gain (SSG) for every user/subchannel pair, which has polynomial complexity. Next, for Markov channels, we outlined a recursive method to update the posterior SSG distributions from the ACK/NAK feedbacks received at each time-slot, and proposed an efficient implementation based on particle filtering. To gauge the performance of our greedy scheme relative to that of the optimal POMDP (which is impossible to implement), we derived a performance upper-bound on POMDP, known as the causal global genie (CGG). Numerical experiments suggest that our greedy scheme achieves a significant fraction of the maximum possible performance gain over fixed-power random user scheduling (FP-RUS), despite its low-complexity implementation. For example, a representative simulation using $N = 32$ OFDMA subchannels, $K = 8$ available users, $\mathsf{SNR}= 10\text{dB}$, and $S = 30$ particles, shows that the sum-goodput of

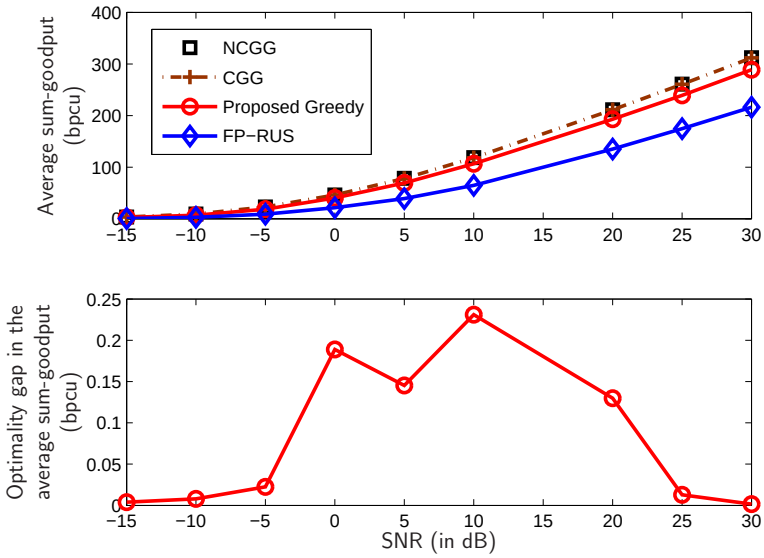

Figure 4.7: The top plot shows the average sum-goodput as a function of SNR. The bottom plot shows the average bound on the optimality gap between the proposed and optimal greedy solutions (given in (4.35)), i.e., the average value of $(\mu^* - \mu_{\min})(X_{\mathsf{con}} - X^*_{\mathsf{tot}}(\mathbf{I}^{\min}, \mu^*))$. In this plot, $N = 32$, $K = 8$, $\alpha = 10^{-3}$, and $S = 30$.

the proposed scheme is 92% of the upper bound and 170% of that attained by the FP-RUS (see Fig. 4.3).

Table 4.4: Particle filtering steps

Let the system begin at time-instant t_0. If $t \in \{t_0, \ldots, t_0 + d - 1\}$:

1. Initialize $\{h_{l,k}^t[i] \ \forall i, k, l\}$ by drawing i.i.d. samples from $\mathcal{CN}(0, \frac{\alpha}{2-\alpha})$.
2. Set the importance weights $\nu_k^{t|t}[i] = \frac{1}{S} \ \forall k, i$.

For any other time-instant t ($\geqslant t_0 + d$):

1. Using the previous samples $\{h_{l,k}^\tau[i] \ \forall l, k, i, \tau : \tau \leqslant t - d\}$, obtain new samples according to the underlying Markov model as follows

$$h_{l,k}^t[i] = (1-\alpha)^d \, h_{l,k}^{t-d}[i] + \alpha \sum_{j=0}^{d-1} (1-\alpha)^j \, y_{l,k}^{t-j}[i], \ \forall i, l, k,$$

   where $y_{l,k}^{t-j}[i]$ is drawn i.i.d from $\mathcal{CN}(0,1)$ for all i, l, k, j.

2. For each user k,

   (a) Using the received feedbacks $\mathbf{f}_k^{t-d}$ and the set of importance weights from time $(t-d-1)$, i.e., $\{\nu_k^{t-d-1|t-d-1}[i] \ \forall i\}$, compute the new set of importance weights at time $(t-d)$ using

$$\nu_k^{t-d|t-d}[i] = \nu_k^{t-d-1|t-d-1}[i] \times p\big(\mathbf{f}_k^{t-d} \,\big|\, \mathbf{h}_k^{t-d} = \mathbf{h}_k^{t-d}[i], \mathbf{I}^{t-d}, \mathbf{P}^{t-d}\big),$$

   for all i, where $p\big(\mathbf{f}_k^{t-d} \,\big|\, \mathbf{h}_k^{t-d} = \mathbf{h}_k^{t-d}[i], \mathbf{I}^{t-d}, \mathbf{P}^{t-d}\big)$ is given by (4.43)-(4.44).

   (b) Normalize the weights via

$$\nu_k^{t-d|t-d}[i] \leftarrow \frac{\nu_k^{t-d|t-d}[i]}{\sum_j \nu_k^{t-d|t-d}[j]} \ \forall i.$$

   (c) Compute the weights for the posterior distribution, $p(\mathbf{h}_k^t \,|\, \mathbb{F}_{-\infty}^{t-d})$ using

$$\nu_k^{t|t-d}[i] = \sum_{j=1}^{S} \nu_k^{t-d|t-d}[j] \, p(\mathbf{h}_k^t = \mathbf{h}_k^t[i] \,\big|\, \mathbf{h}_k^{t-d} = \mathbf{h}_k^{t-d}[j]).$$

   (d) Normalize the weights via

$$\nu_k^{t|t-d}[i] \leftarrow \frac{\nu_k^{t|t-d}[i]}{\sum_j \nu_k^{t|t-d}[j]} \ \forall i.$$

Chapter 5: Large Scale Wireless OFDMA System Design

5.1 Introduction

In Chapters 2-4, we studied some aspects of point-to-point systems and single-transmitter multiple-receiver (OFDMA) systems that are useful in efficient system design. In this chapter, we focus our attention on multiple-transmitter multi-receiver systems.

With the widespread usage of smart phones and an increasing demand for numerous mobile applications, wireless cellular/dense networks have grown significantly in size and complexity. Consequently, the decisions regarding the deployment of transmitters (base-stations, femtocells, picocells etc.), the maximum number of users (subscribers), the amount to be spent on purchasing more bandwidth, and the revenue model to choose have become much more complicated for service providers. Understanding the performance limits of large wireless networks and the optimal balance between the number of serving transmitters, the number of subscribers, the number of antennas used for physical-layer communication, and the amount of available bandwidth to achieve those limits are critical components of the decisions made. Given that the most significant fraction of the performance growth of wireless networks in

the last few decades is associated [66] with cell sizes (that affect interference management schemes) and the amount of available bandwidth, the aforementioned issues become more important.

To answer some of the above questions, we analyze the expected achievable downlink sum-rate in large OFDMA systems as a function of the number of transmitters B, users K, available resource-blocks N, and/or co-located antennas at each transmitter M. Here, a resource block is a collection of subcarriers such that all disjoint sub-collections have associated independently fading channels. Using our analysis, we make the following contributions:

- For a general spatial geometry of transmitters and the end users, we develop novel upper and lower bounds on the average achievable rate as a function of K, B, and N.
- We consider asymptotic scenarios in two networks: dense and regular-extended, in which user nodes have a uniform spatial distribution. We evaluate our bounds for Rayleigh, Nakagami-m, Weibull, and LogNormal fading models along with a truncated path-loss model. To evaluate the bounds, we utilize various results from the *extreme value theory*. We also specify the associated scaling laws in all parameters.
- With the developed bounds we give four design principles for service providers and regulators. In the first scenario, we consider a *dense femtocell network* and develop an asymptotic condition on K, B, and N to guarantee a non-diminishing rate for each user. In the second and third scenarios, we consider extended multicell networks and derive bounds for the choice of user-density K/B in order for the service provider to maximize the revenue per transmitter

and, at the same time, keep the per-user rate above a certain limit. Finally, we consider an extended multicell network and develop asymptotic conditions for K, B, and N to guarantee a minimum return on investment for the service provider.

- For dense and regular-networks, we find a distributed resource allocation scheme that achieves, for a wide choice of $\{K, B, N\}$, a sum-rate scaling equal to that of the upper bound (on achievable sum-rate) that we developed earlier.
- Using the proposed achievability scheme, we show that the achievable sum-rate of peer-to-peer networks increases linearly with the number of coordinating transmit nodes B under fixed power allocation schemes only if $B = O\big(\frac{\log K}{\log \log K}\big)$. Our result extends the result in [67], wherein it was stated that if $B = \Omega(\log K)$, then a linear increase in achievable sum-rate w.r.t. B cannot be achieved. We end our discussion with a note on MISO (Multiple-Input Single-Output) systems, where there are a fixed number of co-located antennas at each transmitter, and obtain a similar distributed resource allocation problem as we found earlier towards achievability of expected achievable sum-rate.

The rest of the chapter is organized as follows. In Section 5.2, we discuss past work. In Section 5.3, we introduce our system model. In Section 5.4, we give general upper and lower bounds on expected achievable sum-rate. We also give, for the cases of dense and regular-extended networks, associated sum-rate scaling laws and four network-design principles. In Section 5.5, we find a deterministic power allocation scheme that governs the proposed distributed achievability scheme, followed by an analysis of peer-to-peer networks. In Section 5.6, we provide details of another achievability

scheme, similar to that developed in Section 5.5, for MISO systems. Finally, we conclude in Section 5.7.

5.2 Related Work

Calculation of achievable performance of wireless networks has been a challenging, and yet an extremely popular problem in the literature. The performance of large networks have been mainly analyzed in the asymptotic regimes and the results have been in the form of scaling laws [67–76] following the seminal work by Gupta and Kumar [68]. Unlike these studies, our main bounds are not asymptotic and we take into account both a distance based power-attenuation law and fading in our model. Scaling laws for channel models incorporating both distance based power-attenuation and fading have been considered in [77–79]. However, these works assume an unbounded path-loss model. Unbounded path loss models affect the asymptotic behavior of the achievable rates significantly. For instance, the capacity scaling law of $\Theta(\log K)$ found in [78,79] arises by exploiting infinite channel-gain of the users close to the transmitter, whereas, without path-loss, the scaling law changes to $\Theta(\log\log K)$. The motivation behind our work is to consider a truncated path-loss model that eliminates the singularity of unbounded path-loss models at zero distance. Further, our analyses take into account the bandwidth and number of transmitters (and/or antennas) in large networks, and provide a distributed achievability scheme that is optimal in scaling sense for a large set of network parameters. To the best of our knowledge, such a study has not been done earlier.

5.3 System Model

We consider a time-slotted OFDMA-based downlink network of B transmitters (or base-stations or femtocells or geographically distributed antennas) and K active users, as shown in Fig. 5.1. The transmitters (TX) lie in a disc of radius $p - R$ ($p > R > 0$), and the users are distributed according to some spatial distribution in a concentric disc of radius p. Under such general settings, Theorem 1 gives bounds on the expected achievable sum-rate of the system. In the sequel, however, we assume, for simplicity, that the transmitter locations are arbitrary and deterministic and the users are uniformly distributed. This model too is quite general and can be applied to several network configurations. For example, it models a *dense network* when transmitter locations are random and the network radius p is fixed. Similarly, it models a *multi-cellular regular extended network* when the transmitters (or base-stations) are located on a regular hexagonal grid with a fixed grid-size, i.e., $p \propto \sqrt{B}$.

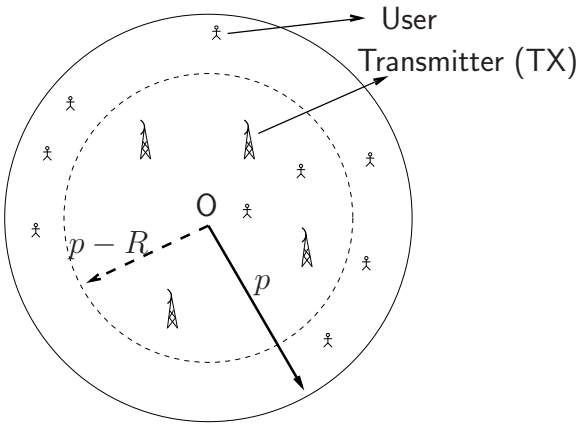

Figure 5.1: OFDMA downlink system with K users and B transmitters. O is assumed to be the origin.

Let us denote the coordinates of TX i ($1 \leq i \leq B$) by (a_i, b_i), and the coordinates of user k ($1 \leq k \leq K$) by (x_k, y_k). Therefore, (a_i, b_i) are known for all i, and (x_k, y_k) is governed by the following probability density function (pdf):

$$f_{(x_k,y_k)}(x,y) = \begin{cases} \frac{1}{\pi p^2} & \text{if } x^2 + y^2 \leq p^2, \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

We now describe the channel model. We assume that the OFDMA subchannels are grouped into N independently-fading resource blocks [80], across which the transmitters (TXs) schedule users for downlink data-transmission. We denote the complex-valued channel gain over resource-block n ($1 \leq n \leq N$) between user k and TX i by $h_{i,k,n}$, and assume that it is composed of the following factors:

$$h_{i,k,n} \;\; = \;\; \beta R_{i,k}^{-\alpha} \nu_{i,k,n}. \tag{5.2}$$

Here, $\beta R_{i,k}^{-\alpha}$ denotes the path-loss attenuation,

$$R_{i,k} = \max\{r_0, \sqrt{(x_k - a_i)^2 + (y_k - b_i)^2}\} \tag{5.3}$$

for constants α, β, r_0 ($\alpha > 1, r_0 < R$), and the fading factor $\nu_{i,k,n}$ is a complex-valued random variable that is i.i.d. across all (i,k,n). Note that r_0 is the truncation parameter that eliminates singularity in the path-loss model. Currently, we keep the distribution of $\nu_{i,k,n}$ general. Specific assumptions on the fading model $\{\nu_{i,k,n}\}$ will be made in subsequent sections. Assuming unit-variance AWGN, the channel Signal-to-Noise Ratio (SNR) between user k and TX i across resource-block n can now be defined as

$$\gamma_{i,k,n} \triangleq |h_{i,k,n}|^2 \;\; = \;\; \beta^2 R_{i,k}^{-2\alpha} |\nu_{i,k,n}|^2. \tag{5.4}$$

We assume that perfect knowledge of the users' channel-SNRs (or gains) from all TXs is available at every transmitter. This can be achieved via a backhaul network

that enables sharing of users' channel-state information[17]. We also assume that the transmitters do not coordinate to send data to a particular user. Therefore, if a user is being served by more than one transmitter, then while decoding the signal from a given TX, it treats the signals from all other TXs as noise. This assumption is restrictive because one may achieve a higher performance by allowing coordination among TXs to send data to users. However, as will be explained after Theorem 1 in Section 5.4, our results and design principles also hold for a class of networks wherein coordination among TXs is allowed.

The maximum achievable sum-rate of our system can now be written as

$$\mathcal{C}_{\mathbf{x},\mathbf{y},\nu}(\mathbf{U},\mathbf{P}) \triangleq \sum_{i=1}^{B}\sum_{n=1}^{N}\log\left(1+\frac{P_{i,n}\,\gamma_{i,U_{i,n},n}}{1+\sum_{j\neq i}P_{j,n}\,\gamma_{j,U_{i,n},n}}\right) \tag{5.5}$$

where $\mathbf{x} := \{x_k \text{ for all } k\}$, $\mathbf{y} := \{y_k \text{ for all } k\}$, $\nu := \{\nu_{i,k,n} \text{ for all } i,k,n\}$, $\mathbf{U} := \{U_{i,n} \text{ for all } i,n\}$, and $\mathbf{P} := \{P_{i,n} \text{ for all } i,n\}$. Here, $U_{i,n}$ is the sum-rate maximizing user scheduled by TX i across resource-block n, and $P_{i,n}$ is the corresponding allocated power. We assume that, in each time-slot, the total power allocated by each TX is upper-bounded by P_{con}. Therefore, $\sum_n P_{i,n} \leq P_{\mathsf{con}}$ for all i. One may also write (5.5) as

$$\mathcal{C}_{\mathbf{x},\mathbf{y},\nu}(\mathbf{U},\mathbf{P}) = \max_{\mathbf{u}\in\mathcal{U},\,\mathbf{p}\in\mathcal{P}}\sum_{i=1}^{B}\sum_{n=1}^{N}\log\left(1+\frac{p_{i,n}\,\gamma_{i,u_{i,n},n}}{1+\sum_{j\neq i}p_{j,n}\,\gamma_{j,u_{i,n},n}}\right), \tag{5.6}$$

where $\mathbf{u} \triangleq \{u_{i,n} \text{ for all } i,n\}$, $\mathbf{p} \triangleq \{p_{i,n} \text{ for all } i,n\}$, and $\{\mathcal{U},\mathcal{P}\}$ are the sets of feasible user allocations and power allocations. In particular,

$$\begin{aligned}\mathcal{U} &\triangleq \big\{\{u_{i,n}\} : 1 \leq u_{i,n} \leq K \text{ for all } i,n\big\} \text{ and}\\ \mathcal{P} &\triangleq \Big\{\{p_{i,n}\} : p_{i,n} \geq 0 \text{ for all } i,n, \text{ and } \sum_n p_{i,n} \leq P_{\mathsf{con}} \text{ for all } i\Big\}.\end{aligned} \tag{5.7}$$

[17]Later, we will propose a distributed resource allocation scheme that does not require any sharing of CSI among the transmitters and its sum-rate scales at the same rate as that of an upper bound on the optimal centralized resource allocation scheme for a wide choice of network parameters.

In the next section, we derive novel upper and lower bounds on the expected value of $\mathcal{C}_{\mathbf{x},\mathbf{y},\nu}(\mathbf{U},\mathbf{P})$ that are later used to determine the scaling laws and develop various network-design guidelines. To state the scaling laws, we use the following notations: for two non-negative functions $f(t)$ and $g(t)$, we write $f(t) = O(g(t))$ if there exists constants $c_1 \in \mathbb{R}^+$ and $r_1 \in \mathbb{R}$ such that $f(t) \leq c_1\, g(t)$ for all $t \geq r_1$. Similarly, we write $f(t) = \Omega(g(t))$ if there exists constants $c_2 \in \mathbb{R}^+$ and $r_2 \in \mathbb{R}$ such that $f(t) \geq c_2\, g(t)$ for all $t \geq r_2$. In other words, $g(t) = O(f(t))$. Finally, we write $f(t) = \Theta(g(t))$ if $f(t) = O(g(t))$ and $f(t) = \Omega(g(t))$.

5.4 Proposed General Bounds on Achievable Sum-Rate

The expected achievable sum-rate of the system can be written, using (5.5), as

$$\mathcal{C}^* = \mathrm{E}\left\{\mathcal{C}_{\mathbf{x},\mathbf{y},\nu}(\mathbf{U},\mathbf{P})\right\}, \tag{5.8}$$

where the expectation is over the SNRs $\{\gamma_{i,k,n}\ \forall\ i,k,n\}$. The following theorem gives upper and lower bounds on (5.8) that depend only on the exogenous channel-SNR process.

Theorem 1 (General bounds)**.** *The expected achievable sum-rate of the system, $\mathcal{C}^*$, can be bounded as:*

$$\begin{aligned}\sum_{i,n} \mathrm{E}\left\{\frac{\log\left(1+P_{\mathsf{con}}\ \max_k \gamma_{i,k,n}\right)}{N+P_{\mathsf{con}}\sum_{j\neq i}\gamma_{j,k,n}}\right\}& \\ \leq \mathcal{C}^* \leq\ \min\Bigg\{\sum_{i,n}\mathrm{E}\left\{\log\left(1+P_{\mathsf{con}}\max_k \gamma_{i,k,n}\right)\right\},& \\ N\sum_i \mathrm{E}\left\{\log\left(1+\frac{P_{\mathsf{con}}}{N}\max_{n,k}\gamma_{i,k,n}\right)\right\}\Bigg\}.&\end{aligned} \tag{5.9}$$

Proof. See Appendix C.1. □

The upper bounds in Theorem 1 are obtained by ignoring interference, and the lower bound is obtained by allocating equal powers $\frac{P_{\mathsf{con}}}{N}$ to every resource-block by

every TX. As mentioned earlier, our bounds, that assume an uncoordinated system, also serve as bounds (up to a constant scaling factor) for the expected max-sum-rate of a class of networks wherein the number of transmitters coordinating to send data to any user on any resource block are bounded. This can be explained using the following argument. Let S transmitters coordinate to send data to user k on resource block n and let $\{\gamma_{1,k,n}, \ldots, \gamma_{S,k,n}\}$ be the corresponding instantaneous exogenous Signal-to-Noise ratios. Then, an upper bound on the sum-rate of those S transmitters across resource block n is $\log\left(1+\left(\sum_{s=1}^{S}\sqrt{P_{s,n}\gamma_{s,k,n}}\right)^2\right)$ [81], where $P_{s,n}$ is the power allocated by transmitter s across resource block n. However, this term is upper bounded by $S\sum_{s=1}^{S}\log(1+P_{s,n}\gamma_{s,k,n})$, which is S times the upper bound on sum-rate obtained by ignoring interference in a completely uncoordinated system (same as that used in Theorem 1). Since S is bounded, the scaling laws for the upper bound and the resulting design principles remain unchanged. The lower bound, on the other hand, assumes no coordination and allocates equal power to every TX and every resource-block. Clearly, by coordinating among transmitters, one can achieve better performance. The above arguments, coupled with the fact that Theorem 1 does not assume any specific channel-fading process or any specific distribution on transmitter and user-locations, make our bounds valid for a wide variety of coordinated and uncoordinated networks. In the next subsection, Section 5.4.1, we evaluate the bounds in Theorem 1 to two classes of networks – *dense* and *regular-extended* – using extreme-value theory, and then provide interesting design principles based on them.

5.4.1 Scaling Laws and Their Applications in Network Design

We first present an analysis of dense networks, followed by an analysis of regular-extended networks. In particular, we use extreme-value theory and Theorem 1 to obtain perfomance bounds and associated scaling laws.

Dense Networks

Dense networks contain a large number of transmitters that are distributed over a fixed area. Typically, such networks occur in dense-urban environments and in dense femtocell deployments. In our system-model, a dense network corresponds to the case in which p is fixed, and K, B, N are allowed to grow. The following two lemmas use extreme-value theory and Theorem 1 to give bounds on the achievable sum-rate of the system for various fading channels.

Theorem 2. *For dense networks with large number of users K and Rayleigh fading channels, i.e., $\nu_{i,k,n} \sim \mathcal{CN}(0,1)\ \forall\ i,k,n$,*

$$\big(\log(1+P_{\mathsf{con}}l_K)+O(1)\big)BNf_{\mathsf{lo}}^{\mathsf{DN}}(r,B,N) \le \mathcal{C}^* \le \big(\log(1+P_{\mathsf{con}}l_K)+O(1)\big)BN, \tag{5.10}$$

where $r>0$ is a constant, $l_K = \beta^2 r_0^{-2\alpha}\log\frac{Kr_0^2}{p^2}$, and $f_{\mathsf{lo}}^{\mathsf{DN}}(r,B,N) = \frac{r^2}{(1+r^2)(N+P_{\mathsf{con}}\beta^2 r_0^{-2\alpha}(1+r)B)}$. The following scaling laws result from (5.10):

$$\begin{aligned}\mathcal{C}^* &= O(BN\log\log K),\ \textit{and}\\ \mathcal{C}^* &= \Omega(\min\{B,N\}\log\log K).\end{aligned} \tag{5.11}$$

Proof. For proof, see Appendix C.2. □

Similar results under different fading models are summarized in the following lemma.

Lemma 9. *If $|\nu_{i,k,n}|$ belongs to either Nakagami-m, Weibull, or LogNormal family of distributions, then, for dense networks, the $\mathcal{C}^*$ satisfies, for large K,*

For Nakagami-(m, w)*:*	$\mathcal{C}^* = O(BN \log\log K)$	$\mathcal{C}^* = \Omega(\min\{B, N\} \log\log K)$
For Weibull (λ, t)*:*	$\mathcal{C}^* = O(BN \log\log^{\frac{2}{t}} K)$	$\mathcal{C}^* = \Omega(\min\{B, N\} \log\log^{\frac{2}{t}} K)$
For LogNormal (a, ω)*:*	$\mathcal{C}^* = O(BN\sqrt{\log K})$	$\mathcal{C}^* = \Omega(\min\{B, N\}\sqrt{\log K})$.

Proof. For proof, see Appendix C.3. □

Based on Theorem 2, we now propose a design principle for large dense networks. In the sequel, we call our system "*scalable*" under a certain condition, if the condition is not violated as the number of users $K \to \infty$.

Principle 1. *In dense femtocell deployments, with the condition that the per-user throughput remains above a certain lower bound, for the system to be scalable, the total number of independent resources* BN *must scale as* $\Omega\big(\frac{K}{\log\log K}\big)$.

We use the dense-network abstraction for a dense femtocell deployment [82] where the service provider wants to maintain a minimum throughput per user. In such cases, a necessary condition that the service provider must satisfy is:

$$\frac{BN\Big(\log(1 + P_{\mathsf{con}}\beta^2 r_0^{-2\alpha} \log\frac{Kr_0^2}{p^2}) + s\Big)}{K} \geq \bar{s}$$

for some $\bar{s} > 0$, where $s = O(1)$. The above equation implies

$$\frac{BN \log\log K}{K} = \Omega(1). \tag{5.12}$$

Therefore, the total number of independent resources BN, i.e., the product of number of transmitters and the number of resource blocks (or bandwidth), must scale no slower than $\frac{K}{\log\log K}$. Otherwise, then the system is not scalable and a minimum per-user throughput requirement cannot be maintained.

Next, we consider another class of networks, namely regular-extended networks, and find performance bounds that motivate the subsequent design guidelines for such networks.

Regular Extended Networks

In extended networks, the area of the network grows with the number of nodes, keeping the number of nodes per unit area fixed. Here we study *regular* extended networks, in which the TXs lie on a regular hexagonal grid as shown in Fig. 5.2. The distance between two neighbouring transmitters is $2R$. Hence, the radius of the network $p = \Theta(R\sqrt{B})$.

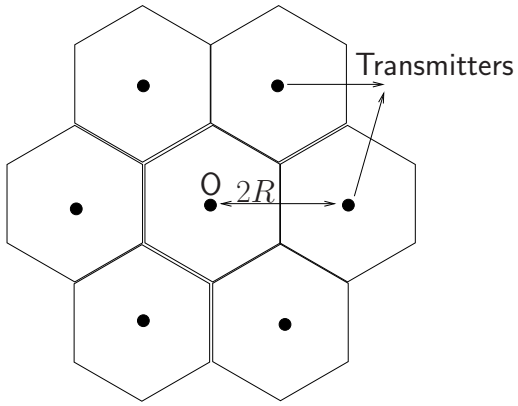

Figure 5.2: A regular extended network setup.

The following two lemmas use Theorem 1 and extreme-value theory to give performance bounds and associated scaling laws for regular extended networks under various fading channels.

Theorem 3. *For regular extended networks with large K and Rayleigh fading channels, i.e., $\nu_{i,k,n} \sim \mathcal{CN}(0,1)$,*

$$\Big(\log(1+P_{\mathsf{con}}l_K)+O(1)\Big)BNf_{\mathsf{lo}}^{\mathsf{EN}}(r,N) \le \mathcal{C}^* \le \Big(\log(1+P_{\mathsf{con}}l_K)+O(1)\Big)BN, \quad (5.13)$$

where $l_K = \beta^2 r_0^{-2\alpha}\log\frac{Kr_0^2}{BR^2}$, $f_{\mathsf{lo}}^{\mathsf{EN}}(r,N) = \frac{(1+r^2)^{-1}r^2}{N+(1+r)c_0}$, and $c_0 = \frac{P_{\mathsf{con}}\beta^2 r_0^{2-2\alpha}}{R^2}\Big(4+\frac{\pi}{\sqrt{3}(2\alpha-2)}\Big)$. The associated scaling laws are:

$$\mathcal{C}^* \;=\; O\Big(BN\log\log\frac{K}{B}\Big), \text{ and } \mathcal{C}^* = \Omega\Big(B\log\log\frac{K}{B}\Big). \quad (5.14)$$

Proof. For proof, see Appendix C.2. □

Lemma 10. *If $|\nu_{i,k,n}|$ belongs to either Nakagami-m, Weibull, or LogNormal family of distributions, then, for regular extended networks, the scaling laws for the upper bounds are:*

For Nakagami-(m,w)*:*	$\mathcal{C}^* = O\Big(BN \log\log \frac{K}{B}\Big)$	$\mathcal{C}^* = \Omega(B \log\log \frac{K}{B})$
For Weibull(λ,t)*:*	$\mathcal{C}^* = O\Big(BN \log\log^{\frac{2}{t}} \frac{K}{B}\Big)$	$\mathcal{C}^* = \Omega\Big(B \log\log^{\frac{2}{t}} \frac{K}{B}\Big)$
For LogNormal(a,ω)*:*	$\mathcal{C}^* = O\Big(BN \sqrt{\log \frac{K}{B}}\Big)$	$\mathcal{C}^* = \Omega\Big(B \sqrt{\log \frac{K}{B}}\Big)$.

Proof. For proof, see Appendix C.3. □

Using Theorem 3, we now propose three design principles.

Principle 2. *In regular extended networks, if a) the users are charged based on the number of bits they download; b) there is a unit cost for each TX installed and a cost c_N for unit resource block incurred by the service provider; c) the return-on-investment must remain above a certain lower bound; then for fixed B, the system is scalable only if $N = O(\log K)$, and for fixed N, the system is scalable only if $B = O(K)$. In addition, if a minimum per-user throughput requirement is also required to be met, then the system is scalable for fixed N only if $B = \Theta(K)$, and not scalable for fixed B.*

Consider the case of a regular extended network with large K. Using the upper bound in Theorem 1 obtained via Jensen's inequality, we have

$$\begin{aligned} \mathcal{C}^* &\leq \left(\log\left(1 + \frac{P_{\mathsf{con}}}{N} l_K + \frac{P_{\mathsf{con}}}{N} \log\log K\right) + O(1)\right) BN \\ &\approx BN \log\left(\frac{P_{\mathsf{con}}}{N} l_K\right), \text{ for large } \frac{P_{\mathsf{con}} l_K}{N}, \end{aligned} \tag{5.15}$$

where $l_K = \beta^2 r_0^{-2\alpha} \log \frac{KNr_0^2}{BR^2}$. For simplicity of analysis, let $P_{\mathsf{con}} = \beta = r_0 = R = 1$ (in their respective SI units). If the service provider wants to maintain a minimum level

of return-on-investment, then it must satisfy

$$\frac{BN}{B+c_N N}\log\left(\frac{1}{N}\log\frac{KN}{B}\right) \;>\; \bar{s}, \tag{5.16}$$

for some $\bar{s}>0$. The above equation implies $N=O(\log K)$ for fixed B, and $B=O(K)$ for fixed N. In addition, if a minimum per-user throughput is also required, then the service provider must satisfy the following equation in addition to (5.16):

$$\frac{BN}{K}\log\left(\frac{1}{N}\log\frac{KN}{B}\right) \;>\; \hat{s}, \tag{5.17}$$

for some $\hat{s}>0$. Equations (5.16) and (5.17) yield that the system is not scalable under fixed B, and for fixed N, the system is scalable only if $B=\Theta(K)$.

Principle 3. *In a large extended multi-cellular network, if the users are charged based on the number of bits they download and there is a unit cost for each TX incurred by the service provider, then there is a finite range of values for the user-density $\frac{K}{B}$ in order to maximize return-on-investment of the service provider while maintaining a minimum per-user throughput.*

Consider a regular extended network with fixed number of resource blocks N. In this case, we have $\mathcal{C}^*=\Theta\big(B\log\log\frac{K}{B}\big)$. Assuming a revenue model wherein the service provider charges per bit provided to the users, the total return on investment of the service provider is proportional to the achievable sum-rate per TX. Therefore, in large scale systems (large K), one must solve:

$$\begin{aligned}&\max_{K,B}\; c\log\left(1+P_{\mathsf{con}}\beta^2 r_0^{-2\alpha}\log\frac{Kr_0^2}{BR^2}\right)\\ &\text{s.t. }\frac{cB\log(1+P_{\mathsf{con}}\beta^2 r_0^{-2\alpha}\log\frac{Kr_0^2}{BR^2})}{K}\geq\bar{s},\end{aligned} \tag{5.18}$$

for some $\bar{s}>0$, where c is a constant bounded according to (5.13)-(5.14). For simplicity, let $\beta=r_0=R=P_{\mathsf{con}}=1$ (in respective SI units). By variable-transformation,

the above problem becomes convex in $\rho \triangleq \frac{K}{B}$. Solving it via dual method, the Karush-Kuhn-Tucker condition is

$$\rho = \frac{(\lambda+1)10}{(1+\log\rho)\lambda}, \tag{5.19}$$

where $\lambda \geq 0$ is the Lagrange multiplier. The plots of LHS and RHS of (5.19) along with the constraint curve are plotted for $\lambda = 0.1, 1, \infty$ in Fig. 5.3. There, the con-

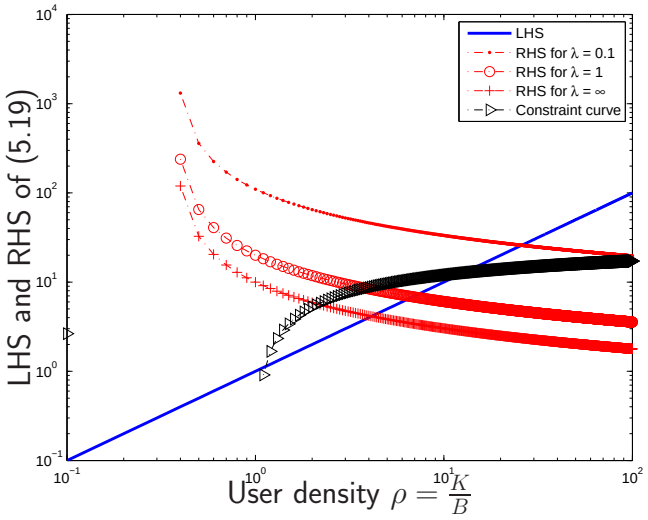

Figure 5.3: LHS and RHS of (5.19) as a function of ρ.

straint curve (see the constraint in (5.18)) is given by $\frac{c}{s}\log(1+\log\rho)$. Note that according to (5.18), the constraint is satisfied only when the constraint curve (in Fig. 5.3) lies above the LHS curve, i.e., when $\rho \in [1.1, 12.7]$. Therefore, the optimal ρ lies in the set $[1.1, 12.7]$. In Figure 5.3, the optimal ρ for a given λ (denoted by $\rho^*(\lambda)$) is the value of ρ at which the LHS and RHS curves intersect for that λ. We observe from the figure that $\rho^*(\lambda)$ decreases with increasing λ. Since $\rho^*(\lambda) = 4.1$ when $\lambda = \infty$, the optimal ρ is greater than or equal to 4.1. Figure 5.4 shows the

variation of $\rho^*(\lambda)$ as a function of λ. From the plot, we observe that $\rho^*(\lambda)$ exists only

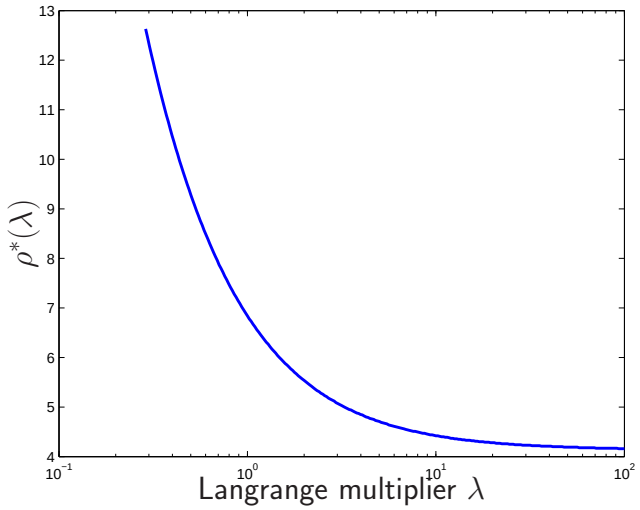

Figure 5.4: Optimal user-density, i.e, $\rho^*(\lambda)$, as a function of λ.

for $\lambda > 0.29$, and satisfies $4.1 \leq \rho^*(\lambda) \leq 12.7$ users/BS. Furthermore, the optimal user-density $\rho^*(\lambda)$ is a strictly-decreasing convex function of the cost associated with violating the per-user throughput constraint, i.e., λ.

Principle 4. *In a large extended multi-cellular network, if the users are charged a fixed amount regardless of the number of bits they download and there is a unit cost for each TX incurred by the service provider, then there is a finite range of values for* $\frac{K}{B}$ *in order to maximize return-on-investement of the service provider while maintaining a minimum per-user throughput.*

Consider a regular extended network with fixed N, similar to that assumed in Principle 3. Here, we assume a revenue model for the service provider wherein the service provider charges each user a fixed amount regardless of the number of bits

the user downloads. Then, the return on investment of the service provider is proportional to the user-density $\rho = \frac{K}{B}$. In large scale systems (large K), the associated optimization problem is:

$$\max_{K,B} s\frac{K}{B} \quad \text{s.t. } \frac{cB\log(1+P_{\mathsf{con}}\beta^2 r_0^{-2\alpha}\log\frac{Kr_0^2}{BR^2})}{K} \geq \bar{s} \tag{5.20}$$

for some constants $c, s, \bar{s} > 0$. Here, s depends on the amount users are charged by the service provider, and c can be bounded according to (5.13)-(5.14). For simplicity of analysis, let $\beta = r_0 = R = P_{\mathsf{con}} = 1$ (in respective SI units). The above problem becomes convex in $\rho \triangleq \frac{K}{B}$. Let the optimal solution be denoted by ρ^*. Now, the constraint in terms of ρ is

$$\frac{\bar{s}}{c} \leq \frac{\log(1+\log\rho)}{\rho}., \tag{5.21}$$

The plot of LHS and RHS of (5.21) as a function of ρ (for $\rho \geq 1$) is plotted in Fig. 5.5. Examining (5.21) and Fig. 5.5, we note that the per-user throughput constraint is satisfied only if $\frac{\bar{s}}{c} \in [0, 0.26]$. Moreover, for a given value of $\frac{\bar{s}}{c}$, the set of feasible ρ lies in a closed set (for which the RHS curve remains above the LHS curve). The maximum value of ρ in this closed set, i.e., the value of ρ at point B in Fig. 5.5, is the one that maximizes the objective in (5.20), i.e., sK/B. Hence, it is the optimal ρ for the given value of $\bar{s}/c$. Let us denote it by $\rho^*(\bar{s}/c)$. Note that $\rho^*(\bar{s}/c) \geq 2.14$ (since point B lies to the right of point A in Fig. 5.5).

If $\bar{s}/c$ is known exactly, then the optimal user-density $\rho^* = \rho^*(\bar{s}/c)$. If not, we can write from (5.13)-(5.14) that $c_{\mathsf{lb}} \leq c \leq c_{\mathsf{ub}}$, for some positive constants $c_{\mathsf{lb}}, c_{\mathsf{ub}}$. Then, $\rho^* \in [\rho^*(\bar{s}/c_{\mathsf{lb}}), \rho^*(\bar{s}/c_{\mathsf{ub}})]$. Moreover, since $\rho^*(\bar{s}/c) \geq 2.14$ for all $\bar{s}/c \in [0, 0.26]$, we have $\rho^*(\bar{s}/c_{\mathsf{ub}}) \geq \rho^*(\bar{s}/c_{\mathsf{lb}}) \geq 2.14$.

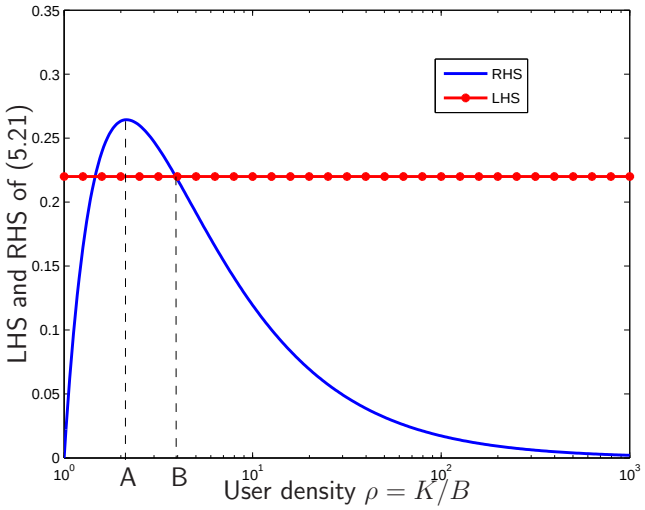

Figure 5.5: LHS and RHS of (5.21) as a function of ρ.

5.5 Maximum Sum-Rate Achievability Scheme

In the previous section, we derived general performance bounds and proposed design principles based on them for two specific types of networks - dense and regular-extended. In this section, we propose a distributed scheme for achievability of max-sum-rate under the above two types of networks. To this end, we construct a tight approximation of $\mathcal{C}^*$ and find a distributed resource allocation scheme that achieves the same sum-rate scaling law as that achieved by $\mathcal{C}^*$ for a large set of network parameters. Let us define an approximation of $\mathcal{C}^*$ as follows:

$$\mathcal{C}^*_{\mathsf{LB}} \triangleq \max_{\mathbf{P}\in\mathcal{P}} \mathrm{E}\left\{ \max_{\mathbf{u}\in\mathcal{U}} \sum_{i=1}^{B}\sum_{n=1}^{N} \log\left(1+\frac{\gamma_{i,u_{i,n},n}\,P_{i,n}}{1+\sum_{j\neq i}\gamma_{j,u_{i,n},n}\,P_{j,n}}\right)\right\}. \tag{5.22}$$

Note that $\mathcal{C}^*_{\mathsf{LB}} \leq \mathcal{C}^*$. To analyze the $\mathcal{C}^*_{\mathsf{LB}}$, we give a novel extreme-value theoretic result in Theorem 4.

Theorem 4. *Let $\{X_1, \ldots, X_T\}$ be i.i.d. random variables with cumulative distribution function (cdf) $F_X(\cdot)$. Then, for any increasing concave function $V(\cdot)$, we have*

$$\big(1 - e^{-S_1}\big) V\big(l_{T/S_1}\big) \;\; \leq \;\; \mathrm{E}\Big\{ V\big(\max_{1 \leq t \leq T} X_t \big) \Big\}. \tag{5.23}$$

Here, $S_1 \in (0, T]$ and $F_X\big(l_{T/S_1}\big) = 1 - \frac{S_1}{T}$.

Proof. See Appendix C.4. □

The following corollary can be stated as a result of the above theorem.

Corollary 1. *Let $\{X_1, \ldots, X_T\}$ be i.i.d. random variables with cdf $F_X(\cdot)$. Then,*

$$\left(1 - \frac{1}{\log T}\right) \log\big(1 + p\, l_{T/\log\log T}\big) \;\; \leq \;\; \mathrm{E}\Big\{ \log\big(1 + p \max_t X_t\big) \Big\}.$$

Furthermore,

$$0.63 \;\; \leq \;\; \frac{\mathrm{E}\Big\{ \log\big(1 + p \max_t X_t\big) \Big\}}{\log\big(1 + p\, l_T\big)}. \tag{5.24}$$

Proof. Substitute $S_1 = \log\log T$, and $V(x) = \log(1 + px)$ in Theorem 4 to prove the first equation. Substitute $S_1 = 0.5$ and $V(x) = \log(1 + px)$ to prove (5.24). □

We will now use the above corollary to find upper and lower bounds on $\mathcal{C}^*_{\mathsf{LB}}$ for dense networks and Rayleigh-fading channels[18] via a class of deterministic optimization problems.

Theorem 5. *Under Rayleigh-fading channels, i.e., $|\nu_{i,k,n}| \sim \mathcal{CN}(0,1)$, let a class of deterministic optimization problems be defined as follows:*

$$\begin{aligned}
&\mathsf{OP}\big(c, h(K)\big) \\
&\quad \triangleq \max_{\mathbf{P} \in \mathcal{P}} \sum_{i=1}^{B} \sum_{n=1}^{N} \log(1 + P_{i,n} x_{i,n}) && (5.25) \\
&\qquad \text{s.t. } x \frac{r_0^2\, h(K)}{p^2} = e^{\frac{x_{i,n}}{\beta^2 r_0^{-2\alpha}}} \prod_{j \neq i} \left(1 + \frac{P_{j,n} x_{i,n}}{c^{2\alpha} r_0^{-2\alpha}}\right) \text{ for all } i, n, && (5.26)
\end{aligned}$$

where $h(\cdot)$ is an increasing function and c is a positive constant. Then, for large K,

$$\big(1 - e^{-S_1}\big) \mathsf{OP}\big(r_0, K/S_1\big) \leq \mathcal{C}^*_{\mathsf{LB}} \leq \left(1 + \frac{\beta^2 r_0^{-2\alpha} u}{\overline{l}(2p, K)}\right) \mathsf{OP}\big(2p, K\big) \tag{5.27}$$

[18] Theorem 5 can be easily extended for Nakagami-m, Weibull, and LogNormal fading channels.

where $S_1 \in (0, K]$, u is the Euler-Mascheroni constant, and $\bar{l}(2p, K)$ is a large number that increases with increasing K. In particular, if $l = \hat{l}(\eta_1, \eta_2)$ be the solution to $\frac{r_0^2 \eta_2}{p^2} = e^{\frac{l}{\beta^2 r_0^{-2\alpha}}} \left(1 + \frac{l P_{\mathsf{con}}}{\eta_1^{2\alpha} r_0^{-2\alpha}}\right)^{B-1}$ for any η_1, η_2, then $\bar{l}(2p, K) \approx \hat{l}(2p, K)$ for large K. Further, $\mathsf{OP}(\cdot, \cdot)$ satisfies

$$1 \leq \frac{\mathsf{OP}\big(c_2, h(K)\big)}{\mathsf{OP}\big(c_1, h(K)\big)} \leq \left(\frac{c_2}{c_1}\right)^{2\alpha} \tag{5.28}$$

for positive constants c_1 and c_2 ($c_1 \leq c_2$).

Proof. See Appendix C.5. □

The above theorem leads to following two corollaries for dense and regular-extended networks.

Corollary 2. *For dense networks (i.e., fixed p) and Rayleigh-fading channels, we have*

$$\left(1 - \frac{1}{\log K}\right) \mathsf{OP}\left(r_0, \frac{K}{\log \log K}\right) \;\leq\; \mathcal{C}_{\mathsf{LB}}^* \leq \left(1 + \frac{\beta^2 r_0^{-2\alpha} u}{\bar{l}(2p, K)}\right) \mathsf{OP}(2p, K) \tag{5.29}$$

and

$$0.63\, \mathsf{OP}(r_0, K) \;\leq\; \mathcal{C}_{\mathsf{LB}}^* \leq \left(\frac{2p}{r_0}\right)^{2\alpha} \left(1 + \frac{\beta^2 r_0^{-2\alpha} u}{\bar{l}(2p, K)}\right) \mathsf{OP}(r_0, K). \tag{5.30}$$

Note that for fixed B and large K, $\bar{l}(2p, K) = \Theta(\log K)$.

Proof. Put $S_1 = \log \log K$ in Theorem 5 to prove (5.29). Put $S_1 = 1$ in (5.27) and use (5.28) to prove (5.30). □

Corollary 3. *For regular extended networks and Rayleigh-fading channels, if $\rho \triangleq K/B$ users are distributed uniformly in each cell and each TX schedules users only within its cell, then*

$$\left(1 - \frac{1}{\log \rho}\right) \mathsf{OP}\left(r_0, \frac{\rho}{\log \log \rho}\right) \leq \mathcal{C}_{\mathsf{LB}}^* \leq \left(1 + \frac{\beta^2 r_0^{-2\alpha} u}{\bar{l}(R\sqrt{3}/2, K)}\right) \mathsf{OP}\left(\frac{R\sqrt{3}}{2}, \rho\right). \tag{5.31}$$

Moreover, we have

$$0.63\, \mathsf{OP}(r_0, \rho) \leq \mathcal{C}_{\mathsf{LB}}^* \leq \left(1 + \frac{\beta^2 r_0^{-2\alpha} u}{\bar{l}(R\sqrt{3}/2, K)}\right) \left(\frac{R\sqrt{3}}{2r_0}\right)^{2\alpha} \mathsf{OP}(r_0, \rho). \tag{5.32}$$

Proof. Note that $p = \Theta(\sqrt{B})$ in this case. Therefore we use, instead of $h(K)$, $h(\rho)$ in Theorem 5 to obtain the above result, where $\rho = \frac{K}{B}$. Also note that $2p$ is replaced by $\frac{R\sqrt{3}}{2}$ since the maximum distance between a user and its serving TX is $\frac{R\sqrt{3}}{2}$. □

The above two corollaries highlight the idea behind the proposed achievability strategy. In particular, we use the lower bounds in (5.30) and (5.32) to give a distributed resource allocation scheme. The steps of the proposed achievability scheme are summarized below.

1. Find the best power allocation $\{P_{i,n}\}$ by solve the LHS of (5.30) for dense networks, or LHS of (5.32) for regular-extended networks. Note that this can be computed offline.

2. For each TX i and resource-block n, schedule the user $k(i,n)$ that satisfies:

$$k(i,n) = \underset{k}{\operatorname{argmax}} \frac{P_{i,n}\gamma_{i,k,n}}{1+\sum_{j\neq i} P_{j,n}\gamma_{j,k,n}}. \tag{5.33}$$

We propose that each user k calculates $\frac{P_{i,n}\gamma_{i,k,n}}{1+\sum_{j\neq i} P_{j,n}\gamma_{j,k,n}}$ for each (i,n) combination and feeds back the value to TX i, thus making the algorithm distributed.

We will now compare low-powered peer-to-peer networks and high-powered single TX systems to give a design principle using on the bounds in Corollary 2.

Principle 5. *The sum-rate of a peer-to-peer network with B transmit nodes (geographically distributed antennas), each transmitting at a fixed power $\bar{P}$ across every resource-block, increases linearly with B only if $B = O\big(\frac{\log K}{\log\log K}\big)$. If $B = \Omega\big(\frac{\log K}{\log\log K}\big)$, then there is no gain with increasing B. Further, the gain obtained by implementing a peer-to-peer network over a high-powered single-TX system (with power $B\bar{P}$ across each resource-block) is*

$$\begin{cases} \Theta(B) & \text{if } B = O\big(\frac{\log K}{\log\log K}\big), \\ \Theta\big(\frac{\log K}{\log\log K}\big) & \text{if } B = \Omega\big(\frac{\log K}{\log\log K}\big) \text{ and } B = O(\log K), \\ \Theta\big(\frac{\log K}{\log B}\big) & \text{if } B = \Omega(\log K). \end{cases} \tag{5.34}$$

In this case, we consider a peer-to-peer networks with B nodes randomly distributed in a circular area of fixed radius p. Assuming fixed power allocation, we

have $P_{i,n} = \bar{P}$ for all i, n. Therefore, from (5.26), we get

$$x_{i,n} \approx \Theta\left(\min\left\{ \beta^2 r_0^{-2\alpha} \log \frac{r_0^2 h(K)}{p^2},\ \frac{1}{\bar{P}} \Big(\frac{c}{r_0}\Big)^{2\alpha\ B-1}\sqrt{\frac{r_0^2}{p^2} h(K)} \right\}\right), \tag{5.35}$$

and $\mathsf{OP}(c, h(K)) = \Theta\big(\sum_{i,n} \log(1+\bar{P}x_{i,n})\big) = \Theta\big(\min\big\{BN \log\log h(K), N \log h(K)\big\}\big)$. Note that in the scenarios of either fixed power allocation schemes or channel coherence time being much smaller than the codeword length, $\mathcal{C}^*_{\mathsf{LB}} = \Theta\big(\mathsf{OP}(c, K)\big)$ is the expected maximum achievable sum-rate $\mathcal{C}^*$. Therefore, using (5.30), we have $h(K) = K$, and the sum-rate capacity under fixed power-allocation scaling as:

$$\mathcal{C}^* = \Theta\big(\min\big\{BN \log\log K, N \log K\big\}\big). \tag{5.36}$$

In other words, if $B = O\big(\frac{\log K}{\log\log K}\big)$, then $\mathcal{C}^* = \Theta(BN \log\log K)$, i.e., we get a linear scaling in sum-rate capacity w.r.t. B. Note that this is also the scaling of the upper bound on sum-rate capacity given in Theorem 2. However, if $B = \Omega\big(\frac{\log K}{\log\log K}\big)$, then $\mathcal{C}^* = \Theta(N \log K)$.

One can also view the above scenario as a multi-antenna system with a single base-station in which all B transmitters are treated as co-located antennas (i.e., $B = M$). Then, comparing our results to those in [67], we note that our results extend the results in [67]. In particular, [67] showed that that linear scaling of sum-rate $\mathcal{C}^*$ w.r.t. number of antennas M holds when $M = \Theta(\log K)$ and does not hold when $M = \Omega(\log K)$. We establish that even if M scales slower than $\log K$, the achievable sum-rate scaling is not linear in M unless $M = O\big(\frac{\log K}{\log\log K}\big)$. Only in the special case of $M = \Theta(\log K)$ is $\mathcal{C}^* = \Theta(N \log K) = \Theta(NM)$. Another way to state the above result is that for a given number of users K (K is large), the achievable sum-rate increases with increasing M only until $M = O\big(\frac{\log K}{\log\log K}\big)$, beyond which the achievable sum-rate stabilizes.

Now, comparing the low-powered peer-to-peer network to a high-powered single-TX system, we assume, for fair comparison, that $P_{1,n} = B\bar{P}\ \forall\ n$. Then, for a high-powered single-TX system, we have $\mathcal{C}^* = \Theta(N \log(B\bar{P} \log K))$. Hence, the gain of peer-to-peer networks over a high-powered single-TX system is given by (5.34).

5.6 A Note on MISO vs SISO Systems

Until now, we discussed systems where either every transmitter had a single antenna or different transmitters were treated as geographically distributed antennas with independent power constraints (i.e., P_{con} at each TX). We wrap up our analysis with a discussion on mutiple antennas at each TX followed by conclusions in Section 5.7.

We use the opportunistic random scheduling scheme proposed in [67], which achieves the sum-rate capacity in the scaling sense for fixed power-allocation schemes. Assume that each TX has M antennas and each user (or, receiver) has a single antenna. Every TX constructs M orthonormal random beams ϕ_m $(M \times 1)$ for $m \in \{1, \ldots, M\}$ using an isotrophic distribution [83]. With some abuse of notation, let us suppose that the signal received by user scheduled by TX i across resource block n using beam m, denoted by $u(i,n,m)$, is given by

$$\begin{aligned} y_{u(i,n,m),n} \;\; &= \;\; \mathbf{H}_{i,u(i,n,m),n}\Big(\phi_m\, x_{i,u(i,n,m),n} + \sum_{m' \neq m} \phi_{m'}\, x_{i,u(i,n,m'),n}\Big) \\ &\quad + \sum_{j \neq i} \sum_{\tilde{m}=1}^{M} \mathbf{H}_{j,u(i,n,\tilde{m}),n}\, \phi_{\tilde{m}}\, x_{j,u(j,n,\tilde{m}),n} + w_{u(i,n,m),n}\,, \end{aligned} \tag{5.37}$$

where $\mathbf{H}_{i,k,n} = \beta R_{i,k}^{-\alpha} \nu_{i,k,n} \in \mathbb{C}^{1\times M}$ is the channel-gain matrix, $\nu_{i,k,n}$ is the $1 \times M$ vector containing i.i.d. complex Gaussian random variables, and $w_{k,n} \sim \mathcal{CN}(0,1)$ is AWGN that is i.i.d. for all (k,n). Abbreviating $\mathrm{E}\{|x_{i,u(i,n,m),n}|^2\}$ by $P_{i,n,m}$, we can

write the SINR corresponding to the combination (i,k,n,m) as:

$$\mathsf{SINR}_{i,k,n,m} \;=\; \frac{P_{i,n,m}\gamma_{i,k,n,m}}{N_0+\sum_{m'\neq m}P_{i,n,m'}\,\gamma_{i,k,n,m}+\sum_{j\neq i}\sum_{\tilde{m}=1}^{M}P_{j,n,\tilde{m}}\,\gamma_{j,k,n,\tilde{m}}}, \quad (5.38)$$

where $\gamma_{i,k,n,m} \triangleq \frac{|\mathbf{H}_{i,k,n}\,\phi_m|^2}{N_0}$ for all (i,k,n,m). Since $\mathbf{H}_{i,k,n}\,\phi_m$ are i.i.d. over all (k,m,n) [67], $\gamma_{i,k,n,m}$ are i.i.d. over (k,m,n). A lower bound on sum-rate capacity, similar to that in (5.22), under opportunistic random beamforming can be written as:

$$\mathcal{C}^{*}_{\mathsf{LB,MISO}} \triangleq \max_{\{P_{i,n,m}\geq 0\ \forall\ i,n,m\}} \mathrm{E}\left\{\max_{\{u(i,n,m)\}}\sum_{i=1}^{B}\sum_{n=1}^{N}\sum_{m=1}^{M}\log\left(1+\mathsf{SINR}_{i,u(i,n,m),n,m}\right)\right\} \quad (5.39)$$

$$\text{s.t.}\ \sum_{n,m}P_{i,n,m}\leq P_{\mathsf{con}}\ \text{for all}\ i. \quad (5.40)$$

The above optimization problem is similar to that in (5.22) with BM transmitters. Therefore, repeating the analysis in (5.22)-(5.30) under dense networks for the problem in (5.39)-(5.40), we get

$$\left(1-\frac{1}{\log K}\right)\mathsf{OP}_{\mathsf{MISO}}\left(r_0,\frac{K}{\log\log K}\right) \;\leq\; \mathcal{C}^{*}_{\mathsf{LB,MISO}} \quad (5.41)$$

$$\leq\; \left(1+O\left(\frac{1}{\log K}\right)\right)\mathsf{OP}_{\mathsf{MISO}}(2p,K),$$

where

$$\mathsf{OP}_{\mathsf{MISO}}(c,h(K)) \triangleq \max_{\{P_{i,n,m}\geq 0\ \forall n,m,n\}}\sum_{i=1}^{B}\sum_{n=1}^{N}\sum_{m=1}^{M}\log(1+P_{i,n,m}\,x_{i,k,n,m}) \quad (5.42)$$

$$\text{s.t.}\ \sum_{m,n}P_{i,n,m}\leq P_{\mathsf{con}}\ \forall\ i,\ \text{and for all}\ (i,m),$$

$$\left(1+\frac{P_{i,n,m}x_{i,k,n,m}}{c^{2\alpha}r_0^{-2\alpha}}\right)\frac{r_0^2h(K)}{p^2}=e^{\frac{x_{i,k,n,m}}{\beta^2r_0^{-2\alpha}}}\prod_{j}\prod_{\tilde{m}=1}^{M}\left(1+\frac{P_{j,n,\tilde{m}}x_{i,k,n,\tilde{m}}}{c^{2\alpha}r_0^{-2\alpha}}\right).$$

Furthermore,

$$0.63 \leq \frac{\mathcal{C}^{*}_{\mathsf{LB,MISO}}}{\mathsf{OP}_{\mathsf{MISO}}\big(r_0, K\big)} \leq \left(\frac{2p}{r_0}\right)^{2\alpha} + O\bigg(\frac{1}{\log K}\bigg). \tag{5.43}$$

5.7 Conclusion

In this paper, we developed bounds for the downlink sum-rate capacity in large OFDMA based networks and derived the associated scaling laws with respect to number of users K, number of transmitters B, and number of resource-blocks N. Our bounds hold for a general spatial distribution of transmitters, a truncated path-loss model, and a general channel-fading model. We evaluated the bounds in *dense* and *extended* networks in which tranmitter nodes are distributed uniformly for Rayleigh, Nakagami-m, Weibull, and LogNormal fading models. Using these results, we developed four design principles for service providers and regulators to achieve QoS provisioning along with system scalability. According to the first principle, in dense-femtocell deployments, for a minimum per-user throughput requirement, we showed that then the system is scalable only if BN scales as $\Omega\big(\frac{K}{\log\log K}\big)$. In the second and third principles, we considered different pricing policies and showed that the user density must be kept within a finite range of values in order to maximize the return-on-investment, while maintaining a minimum per-user rate. In the fourth principle, we also considered the cost of bandwidth to the service provider along with the cost of the transmitters and showed that, for fixed B, the system is scalable only if $N = O(\log K)$, and for fixed N, the system is scalable only if $B = O(K)$. Thereafter, towards developing an achievability scheme, we proposed a deterministic distributed resource allocation scheme and developed more design principles. In particular, we showed that the sum-rate capacity of a peer-to-peer network with B transmitters (or,

a MISO system with B transmitter antennas at a single transmit node) increases with B only when $B = O\big(\frac{\log K}{\log \log K}\big)$.

Chapter 6: Conclusions and Future Work

In this dissertation, we studied downlink resource allocation strategies for point-to-point systems, single-cell OFDMA (Orthogonal Frequency Division Multiple Access) systems, and multi-cell OFDMA systems and proposed guidelines for service providers to design efficient wireless communication systems. First, for point-to-point systems, we proposed greedy rate-adaptation schemes based on ACK/NAK (Acknowledgement/Negative Acknowledgement) feedback for continuous-state channels. We showed that our greedy rate adaptation scheme performs significantly better than the fixed rate scheme and is close to an upper bound on the optimal POMDP-based rate-adaptation scheme, especially under slow-fading channels. Second, for single-cell OFDMA downlink systems, we proposed simultaneous user-scheduling and resource (power and rate) allocation algorithms under imperfect channel-state information. In cases where subchannel-sharing among users is allowed, we propose an optimal algorithm and for cases in which one subchannel is assigned to one user only, we proposed an algorithm that was near-optimal. Our algorithm is faster than the traditional subgradient based/golden-section based algorithms. Further, we gave theoretical performance guarantees as a function of number of iterations of the algorithm. Finally, we considered large multi-cellular OFDMA-based networks and proposed performance bounds as a function of the number of users K, the number of base-stations B, and

the number of resource-blocks N. In particular, we derived novel upper and lower bounds on the achievable sum-rate for a general spatial geometry of transmitters (or, base-stations), a truncated path loss model, and a variety of fading models (Rayleigh, Nakagami-m, Weibull, and LogNormal). We also derived the associated scaling laws and developed design principles for service providers, along with some guidelines for the regulators, in order to achieve provisioning of various QoS (Quality of Service) guarantees for the end users and, at the same time, maximize revenue for the service providers. Furthermore, we provided a scheme that achieves the same sum-rate scaling as that of the optimal resource allocation scheme.

Some future work directions in which we wish to continue our work are as follows:

1. Considering imperfect channel-state information (CSI) in multi-transmitter systems and developing distributed scheduling and resource allocation schemes (similar to those proposed in Chapter 5) that achieve the same sum-rate scaling as that of the optimal resource allocation scheme remains a topic of future work. The motivation for this line of work comes from the question: How does the accuracy of CSI affect the sum-rate of a communication system? Another relevant question that needs to be answered is: Under a given pricing scheme, what trade-offs exist between accuracy of CSI and the accumulated revenue?

2. Our analysis in Chapter 5 was based on transmitters that did not cooperate to send data to a particular receiver. Similar analyses can be done for systems that allow transmitter cooperation and beamforming techniques at the transmitters/receivers. This may possibly lead to the answer to questions such as: When does beamforming fail to provide significant gain in the sum-rate of large

dense networks? The motivation behind this idea is that, in large dense networks, due to availability of large number of transmitters within a fixed area, the optimal resource allocation scheme might force some transmitters to not transmit at all. This will reduce the interference caused to other transmitters' signals received at user-terminals. However, it will also limit the beamforming gain. Note that, with large number of transmitters in a dense network (with fixed network size), there will always exist transmitters that are close to a given user and observe strong channel conditions between themselves and that user. Thus, user scheduling alone may perform close to beamforming.

3. Extensions can also be made to OFDMA systems with multiple antennas at the transmitters and receivers, or large MIMO-OFDMA (Multiple Input Multiple Output OFDMA) systems. Related results involving multiple antennas at transmitters and single antenna at receivers have been discussed in Chapter 5. However, a more general analysis for MIMO-OFDMA systems remains a topic of future work.

4. Another potential problem involves resource allocation under a combination of both uplink and downlink communication. Consider the case where only two users (say, A and B) communicate via a communication link. The transmitting user (say, A) sends data to its serving base-station (say BS-A), which then sends the data to another base-station (say, BS-B) that serves receiving user B. The amount of data that can be sent on this communication link is limited by the channel conditions in uplink channel (from A to BS-A), downlink channel (from BS-B to B), and the rate of incoming data at the transmitting user A. Thus,

allocation of resources needs to be done so that the sum-rate/throughput of this system is maximized, where the sum-rate/throughput should take into account all the aforementioned factors, i.e., arrival rate of data and uplink & downlink channel-conditions.

Appendix A: Proofs in Chapter 2

Here, we derive the expression for $p(\gamma_t \mid \gamma_{t-nd})$ given in (2.44). Let $g_{t,R}$ and $g_{t,I}$ be the real and imaginary parts of channel gain, g_t. Also let $g_{t-nd} = |g_{t-nd}|e^{j\theta}$ for $\theta \sim \mathrm{U}(0, 2\pi)$. Then

$$p(\gamma_t \mid \gamma_{t-nd}) = \int_0^{2\pi} p(\gamma_t \mid \gamma_{t-nd}, \theta)\, p(\theta) d\theta. \tag{A.1}$$

We first find $p(|g_t| \mid \gamma_{t-nd}, \theta)$ in order to evaluate $p(|g_t| \mid \gamma_{t-nd})$. Since

$$g_t = (1-\alpha)^{nd}|g_{t-nd}|e^{j\theta} + Z \tag{A.2}$$

for $Z = \alpha \sum_{i=0}^{nd-1}(1-\alpha)^j w_{t-j}$ and $|g_t| = \sqrt{\frac{\gamma_t}{K}}$, then, conditional on the pair (γ_{t-nd}, θ), the random variables $g_{t,R}$ and $g_{t,I}$ are both Gaussian with mean

$$\mathrm{E}\{g_{t,R} \mid \gamma_{t-nd}, \theta\} = (1-\alpha)^{nd}\sqrt{\frac{\gamma_{t-nd}}{K}}\cos\theta \tag{A.3}$$

and

$$\mathrm{E}\{g_{t,I} \mid \gamma_{t-nd}, \theta\} = (1-\alpha)^{nd}\sqrt{\frac{\gamma_{t-nd}}{K}}\sin\theta, \tag{A.4}$$

respectively, and variance $\sigma_Z^2 = \mathrm{E}\{Z^2\}$. Thus conditional on (γ_{t-nd}, θ), the random variable $|g_t| = g_{t,R}^2 + g_{t,I}^2$ is Rician [1, p. 78]:

$$p(|g_t| \mid \gamma_{t-nd}, \theta) = \frac{|g_t|}{\sigma_Z^2}\exp\left(\frac{-\left(|g_t|^2 + (1-\alpha)^{2nd}\frac{\gamma_{t-nd}}{K}\right)}{2\sigma_Z^2}\right) \times I_0\left(\frac{|g_t|(1-\alpha)^{nd}\sqrt{\frac{\gamma_{t-nd}}{K}}}{\sigma_Z^2}\right). \tag{A.5}$$

One can see that, given γ_{t-nd}, the random variable $|g_t|$ is independent of θ. Since $\gamma_t = K|g_t|^2$, we have

$$\begin{aligned} p(\gamma_t \mid \gamma_{t-nd}) &= \frac{1}{2K\sigma_Z^2} \exp\left(\frac{-\left(\frac{\gamma_t}{K} + (1-\alpha)^{2nd}\frac{\gamma_{t-nd}}{K}\right)}{2\sigma_Z^2} \right) \\ &\quad \times I_0\left(\frac{(1-\alpha)^{nd}\sqrt{\gamma_t\gamma_{t-nd}}}{K\sigma_Z^2} \right). \end{aligned} \tag{A.6}$$

Hence combining (A.1) and (A.6), we get

$$\begin{aligned} p(\gamma_t \mid \gamma_{t-nd}) &= \frac{1}{2K\sigma_Z^2} \exp\left(\frac{-\left(\gamma_t + (1-\alpha)^{2nd}\gamma_{t-nd}\right)}{2K\sigma_Z^2} \right) \\ &\quad \times I_0\left(\frac{(1-\alpha)^{nd}\sqrt{\gamma_t\gamma_{t-nd}}}{K\sigma_Z^2} \right). \end{aligned} \tag{A.7}$$

Finally, plugging $\sigma_Z^2 = \frac{\alpha}{2-\alpha}\left(1-(1-\alpha)^{2nd}\right)$ into (A.7) yields (2.44).

Appendix B: Proofs in Chapter 3

B.1 Proof for convexity of CSRA problem

In this proof, we will show that the primal objective function in the CSRA problem given by

$$\sum_{n,k,m} I_{n,k,m} F_{n,k,m}(I_{n,k,m}, x_{n,k,m}), \tag{B.1}$$

where $F_{n,k,m}(I_{n,k,m}, x_{n,k,m})$ is defined in (3.4), is a convex function of $\mathbf{I}$ ($\in \mathcal{I}_{\mathsf{CSRA}}$) and $\mathbf{x}$. For this, first we consider the case where $I_{n,k,m} > 0 \ \forall n, k, m$. Then, we have the differentiable function given by

$$I_{n,k,m} F_{n,k,m}(I_{n,k,m}, x_{n,k,m}) = -I_{n,k,m}\, \mathrm{E}\Big\{ \underbrace{U_{n,k,m}\big((1 - a_m e^{-b_m x_{n,k,m}\gamma_{n,k}/I_{n,k,m}}) r_m\big)}_{\tilde{U}_{n,k,m}} \Big\}, \tag{B.2}$$

where the expectation is over $\gamma_{n,k}$. We show that the Hessian of this function with respect to $I_{n,k,m}$ and $x_{n,k,m}$ is positive semi-definite. For this, first we write the gradient of $-I_{n,k,m}\tilde{U}_{n,k,m}$ w.r.t. $I_{n,k,m}$ and $x_{n,k,m}$ as

$$\begin{bmatrix} -\tilde{U}_{n,k,m} + \frac{a_m b_m r_m x_{n,k,m}}{I_{n,k,m}} \tilde{U}'_{n,k,m} \gamma_{n,k}\, e^{-b_m x_{n,k,m}\gamma_{n,k}/I_{n,k,m}} \\ -a_m b_m r_m \tilde{U}'_{n,k,m} \gamma_{n,k}\, e^{-b_m x_{n,k,m}\gamma_{n,k}/I_{n,k,m}} \end{bmatrix}^{\mathsf{T}}, \tag{B.3}$$

where T denotes the transpose of a matrix and $'$ denotes the derivative. For simplicity, let us denote $a_m b_m r_m \gamma_{n,k} e^{-b_m x_{n,k,m}\gamma_{n,k}/I_{n,k,m}}$ as $W_{n,k,m}$. Therefore, the gradient can be

written as:

$$\left[\begin{array}{c} -\tilde{U}_{n,k,m} + \frac{x_{n,k,m}}{I_{n,k,m}} \tilde{U}'_{n,k,m} W_{n,k,m} \\ -\tilde{U}'_{n,k,m} W_{n,k,m} \end{array} \right], \tag{B.4}$$

where

$$\begin{aligned} \frac{\partial \tilde{U}_{n,k,m}}{\partial I_{n,k,m}} &= -\frac{x_{n,k,m}}{I^2_{n,k,m}} \tilde{U}'_{n,k,m} W_{n,k,m}, \\ \frac{\partial \tilde{U}_{n,k,m}}{\partial x_{n,k,m}} &= \frac{\tilde{U}'_{n,k,m} W_{n,k,m}}{I_{n,k,m}}, \\ \frac{\partial W_{n,k,m}}{\partial I_{n,k,m}} &= \frac{b_m x_{n,k,m} \gamma_{n,k}}{I^2_{n,k,m}} W_{n,k,m}, \text{ and} \\ \frac{\partial W_{n,k,m}}{\partial x_{n,k,m}} &= -\frac{b_m \gamma_{n,k}}{I_{n,k,m}} W_{n,k,m}. \end{aligned} \tag{B.5}$$

Using the gradient, the Hessian of $-I_{n,k,m}\tilde{U}_{n,k,m}$ is given as:

$$\left[\begin{array}{cc} \frac{x^2_{n,k,m} W_{n,k,m}}{I^3_{n,k,m}} \left(-\tilde{U}''_{n,k,m} W_{n,k,m} + b_m \gamma_{n,k} \tilde{U}'_{n,k,m} \right) & \frac{x_{n,k,m} W_{n,k,m}}{I^2_{n,k,m}} \left(\tilde{U}''_{n,k,m} W_{n,k,m} - b_m \gamma_{n,k} \tilde{U}'_{n,k,m} \right) \\ \frac{x_{n,k,m} W_{n,k,m}}{I^2_{n,k,m}} \left(\tilde{U}''_{n,k,m} W_{n,k,m} - b_m \gamma_{n,k} \tilde{U}'_{n,k,m} \right) & \frac{W_{n,k,m}}{I_{n,k,m}} \left(-\tilde{U}''_{n,k,m} W_{n,k,m} + b_m \gamma_{n,k} \tilde{U}'_{n,k,m} \right). \end{array} \right]$$

$$= \frac{W_{n,k,m}}{I_{n,k,m}} \left(-\tilde{U}''_{n,k,m} W_{n,k,m} + b_m \gamma_{n,k} \tilde{U}'_{n,k,m} \right) \left[\begin{array}{cc} \frac{x^2_{n,k,m}}{I^2_{n,k,m}} & -\frac{x_{n,k,m}}{I_{n,k,m}} \\ -\frac{x_{n,k,m}}{I_{n,k,m}} & 1. \end{array} \right].$$

The above matrix is positive semi-definite. Therefore, $I_{n,k,m} F_{n,k,m}(I_{n,k,m}, x_{n,k,m})$ is convex in $I_{n,k,m}$ and $x_{n,k,m}$ whenever $I_{n,k,m} > 0$.

Now, if $I_{n,k,m} = 0$, in order to show convexity, we consider 2 points in the domain of CSRA problem as $(I^{(1)}_{n,k,m}, x^{(1)}_{n,k,m})$ and $(I^{(2)}_{n,k,m}, x^{(2)}_{n,k,m})$. Then, for a given $\lambda \in [0, 1]$, we should have

$$\lambda I^{(1)}_{n,k,m} F_{n,k,m}\Big(I^{(1)}_{n,k,m}, x^{(1)}_{n,k,m}\Big) + (1-\lambda) I^{(2)}_{n,k,m} F_{n,k,m}\Big(I^{(2)}_{n,k,m}, x^{(2)}_{n,k,m}\Big) \geq$$

$$\Big[\lambda I^{(1)}_{n,k,m} + (1-\lambda) I^{(2)}_{n,k,m}\Big] F_{n,k,m}\Big(\lambda I^{(1)}_{n,k,m} + (1-\lambda) I^{(2)}_{n,k,m}, \lambda x^{(1)}_{n,k,m} + (1-\lambda) x^{(2)}_{n,k,m}\Big) \tag{B.6}$$

In this case, without loss of generality, let us suppose that for a particular (n, k, m), $I^1_{n,k,m} = 0$. Then, the above inequality is equivalent to

$$I^{(2)}_{n,k,m} F_{n,k,m}\Big(I^{(2)}_{n,k,m}, x^{(2)}_{n,k,m}\Big) \geq I^{(2)}_{n,k,m} F_{n,k,m}\Big((1-\lambda) I^{(2)}_{n,k,m}, \lambda x^{(1)}_{n,k,m} + (1-\lambda) x^{(2)}_{n,k,m}\Big). \tag{B.7}$$

If $I^{(2)}_{n,k,m} = 0$, the above inequality is satisfied with equality. If not, it becomes equivalent to

$$F_{n,k,m}\Big(I^{(2)}_{n,k,m}, x^{(2)}_{n,k,m}\Big) \geq F_{n,k,m}\Big((1-\lambda)I^{(2)}_{n,k,m}, \lambda x^{(1)}_{n,k,m} + (1-\lambda)x^{(2)}_{n,k,m}\Big) \tag{B.8}$$

or

$$\begin{aligned} &- \mathrm{E}\left\{U_{n,k,m}\left(\left(1 - a_m \exp\left(-b_m \gamma_{n,k} \frac{x^{(2)}_{n,k,m}}{I^{(2)}_{n,k,m}}\right)\right) r_m\right)\right\} \geq \\ - \mathrm{E}&\left\{U_{n,k,m}\left(\left(1 - a_m \exp\left(-b_m \gamma_{n,k}\left[\frac{x^{(2)}_{n,k,m}}{I^{(2)}_{n,k,m}} + \frac{\lambda x^{(1)}_{n,k,m}}{(1-\lambda)I^{(2)}_{n,k,m}}\right]\right)\right) r_m\right)\right\} \end{aligned} \tag{B.9}$$

In the above equation, the argument of $U_{n,k,m}(\cdot)$ in the RHS is greater, if not equal, than in the LHS. Since $U_{n,k,m}(\cdot)$ is an increasing concave function, $-U_{n,k,m}(\cdot)$ is a decreasing convex function. Therefore, (B.9) is true.

The above proof shows that $-I_{n,k,m}\tilde{U}_{n,k,m}$ (and, by extension $I_{n,k,m}\, F_{n,k,m}(I_{n,k,m}, x_{n,k,m})$) is convex in $I_{n,k,m}$ and $x_{n,k,m}$. Therefore, it is a convex function of $\mathbf{I}$ and $\mathbf{x}$. Since the primal objective function of the CSRA problem, i.e.,

$$\sum_{n,k,m} I_{n,k,m}\, F_{n,k,m}(I_{n,k,m}, x_{n,k,m})$$

is a sum of functions that are convex in $\mathbf{I}$ and $\mathbf{x}$, it is also convex in $\mathbf{I}$ and $\mathbf{x}$.

B.2 Proof of Lemma 2

Say that $\mu_1 < \mu_2$, where $\mu_1, \mu_2 \in [\mu_{\mathsf{min}}, \mu_{\mathsf{max}}]$. Recall from (3.6) that, after fixing μ, the minimization problem reduces to

$$\begin{aligned} &L(\mu, \mathbf{I}^*(\mu), \mathbf{x}^*(\mu, \mathbf{I}^*(\mu))) \\ &= \min_{\substack{\{\mathbf{x}\succeq 0\} \\ \mathbf{I}\in\mathcal{I}_{\mathsf{CSRA}}}} L(\mu, \mathbf{I}, \mathbf{x}) \\ &= \min_{\substack{\{\mathbf{x}\succeq 0\} \\ \mathbf{I}\in\mathcal{I}_{\mathsf{CSRA}}}} \Big(\sum_{n,k,m} x_{n,k,m} - P_{\mathsf{con}}\Big)\mu + \sum_{n,k,m} I_{n,k,m} F_{n,k,m}(I_{n,k,m}, x_{n,k,m}). \end{aligned} \tag{B.10}$$

Suppose that $\mu = \mu_1$, so that (from (B.10)) the optimal values of $\mathbf{I}^*(\mu)$ and $\mathbf{x}^*(\mu, \mathbf{I}^*(\mu))$ are $\mathbf{I}^*(\mu_1)$ and $\mathbf{x}^*(\mu_1, \mathbf{I}^*(\mu_1))$, respectively. Then, we know

$$L(\mu_1, \mathbf{I}^*(\mu_1), \mathbf{x}^*(\mu_1, \mathbf{I}^*(\mu_1))) \;\leq\; L(\mu_1, \mathbf{I}^*(\mu_2), \mathbf{x}^*(\mu_2, \mathbf{I}^*(\mu_2))), \tag{B.11}$$

since $\mathbf{I}^*(\mu_2)$ and $\mathbf{x}^*(\mu_2, \mathbf{I}^*(\mu_2))$ are suboptimal values of $\mathbf{I}^*(\mu)$ and $\mathbf{x}^*(\mu, \mathbf{I}^*(\mu))$. Similarly, we can say that

$$L(\mu_2, \mathbf{I}^*(\mu_2), \mathbf{x}^*(\mu_2, \mathbf{I}^*(\mu_2))) \;\leq\; L(\mu_2, \mathbf{I}^*(\mu_1), \mathbf{x}^*(\mu_1, \mathbf{I}^*(\mu_1))). \tag{B.12}$$

Abbreviating

$$I^*_{n,k,m}(\mu) F_{n,k,m}\big(\mathbf{I}^*(\mu), \mathbf{x}^*(\mu, \mathbf{I}^*(\mu))\big)$$

by $G^*_{n,k,m}(\mu)$, and using (B.10), we then have

$$\begin{aligned} &-\mu_1 P_{\mathsf{con}} + \sum_{n,k,m} \Big(x^*_{n,k,m}(\mu_1, \mathbf{I}^*(\mu_1))\mu_1 + G^*_{n,k,m}(\mu_1) \Big) \\ &\quad\leq\; -\mu_1 P_{\mathsf{con}} + \sum_{n,k,m} \Big(x^*_{n,k,m}(\mu_2, \mathbf{I}^*(\mu_2))\mu_1 + G^*_{n,k,m}(\mu_2) \Big), \end{aligned} \tag{B.13}$$

and

$$\begin{aligned} &-\mu_2 P_{\mathsf{con}} + \sum_{n,k,m} \Big(x^*_{n,k,m}(\mu_2, \mathbf{I}^*(\mu_2))\mu_2 + G^*_{n,k,m}(\mu_2) \Big) \\ &\quad\leq\; -\mu_2 P_{\mathsf{con}} + \sum_{n,k,m} \Big(x^*_{n,k,m}(\mu_1, \mathbf{I}^*(\mu_1))\mu_2 + G^*_{n,k,m}(\mu_1) \Big), \end{aligned} \tag{B.14}$$

Adding (B.13) and (B.14), we get

$$(\mu_1 - \mu_2)\Big(\sum_{n,k,m} x^*_{n,k,m}(\mu_1, \mathbf{I}^*(\mu_1)) - x^*_{n,k,m}(\mu_2, \mathbf{I}^*(\mu_2)) \Big) \leq 0. \tag{B.15}$$

Since $\mu_1 < \mu_2$, we have

$$X^*_{\mathsf{tot}}(\mu_1) \geq X^*_{\mathsf{tot}}(\mu_2). \tag{B.16}$$

Therefore, $X^*_{\mathsf{tot}}(\mu)$ is monotonically decreasing in μ.

B.3 Proof of Lemma 3

To compare the utilities obtained by the proposed CSRA algorithm and the exact CSRA solution, we compare the Lagrangian values achieved by the two solutions. Recall $\mu^* \in [\underline{\mu}, \bar{\mu}] \subset [\mu_{\mathsf{min}}, \mu_{\mathsf{max}}]$. Therefore,

$$\begin{aligned} L(\mu^*, \mathbf{I}^*(\mu^*), \mathbf{x}^*(\mu^*, \mathbf{I}^*(\mu^*))) - L(\underline{\mu}, \mathbf{I}^*(\underline{\mu}), \mathbf{x}^*(\underline{\mu}, \mathbf{I}^*(\underline{\mu}))) &\geq 0, \text{ and} \\ L(\mu^*, \mathbf{I}^*(\mu^*), \mathbf{x}^*(\mu^*, \mathbf{I}^*(\mu^*))) - L(\bar{\mu}, \mathbf{I}^*(\bar{\mu}), \mathbf{x}^*(\bar{\mu}, \mathbf{I}^*(\bar{\mu}))) &\geq 0. \end{aligned} \tag{B.17}$$

The solution of the proposed CSRA algorithm allocates resources such that the sum-power constraint is satisfied while achieving a Lagrangian value of

$$\hat{L}_{\mathsf{CSRA}} \triangleq \lambda L(\bar{\mu}, \mathbf{I}^*(\bar{\mu}), \mathbf{x}^*(\bar{\mu}, \mathbf{I}^*(\bar{\mu}))) + (1-\lambda) L(\underline{\mu}, \mathbf{I}^*(\underline{\mu}), \mathbf{x}^*(\underline{\mu}, \mathbf{I}^*(\underline{\mu}))).$$

For any μ, notice that $L(\mu, \mathbf{I}^*(\mu), \mathbf{x}^*(\mu, \mathbf{I}^*(\mu))) = -U^*(\mu) + (X^*_{\mathsf{tot}}(\mu) - P_{\mathsf{con}})\mu$, where $U^*(\mu)$ is the total utility achieved due to optimal power allocation at that μ. Since the resource allocation obtained by the proposed CSRA algorithm and the exact CSRA solution satisfy the sum-power constraint with equality, we have

$$U^*_{\mathsf{CSRA}} = -L(\mu^*, \mathbf{I}^*(\mu^*), \mathbf{x}^*(\mu^*, \mathbf{I}^*(\mu^*))), \text{ and} \tag{B.18}$$

$$\begin{aligned} \hat{L}_{\mathsf{CSRA}} &= -\hat{U}_{\mathsf{CSRA}}(\underline{\mu}, \bar{\mu}) + (X^*_{\mathsf{tot}}(\bar{\mu}) - P_{\mathsf{con}})\lambda\bar{\mu} + (X^*_{\mathsf{tot}}(\underline{\mu}) - P_{\mathsf{con}})(1-\lambda)\underline{\mu} \\ &= -\hat{U}_{\mathsf{CSRA}}(\underline{\mu}, \bar{\mu}) + (X^*_{\mathsf{tot}}(\bar{\mu}) - P_{\mathsf{con}})(\bar{\mu} - \underline{\mu})\lambda. \end{aligned} \tag{B.19}$$

Equation (B.19) holds since $\lambda X^*_{\mathsf{tot}}(\bar{\mu}) + (1-\lambda) X^*_{\mathsf{tot}}(\underline{\mu}) = P_{\mathsf{con}}$. From (B.18) and (B.19), we get

$$\begin{aligned} 0 \leq U^*_{\mathsf{CSRA}} - \hat{U}_{\mathsf{CSRA}}(\underline{\mu}, \bar{\mu}) &= -L(\mu^*, \mathbf{I}^*(\mu^*), \mathbf{x}^*(\mu^*, \mathbf{I}^*(\mu^*))) + \hat{L}_{\mathsf{CSRA}} \\ &\quad - (X^*_{\mathsf{tot}}(\bar{\mu}) - P_{\mathsf{con}})(\bar{\mu} - \underline{\mu})\lambda. \end{aligned} \tag{B.20}$$

From the above equation and (B.17), we have

$$0 \;\leq\; U^*_{\mathsf{CSRA}} - \hat{U}_{\mathsf{CSRA}}(\underline{\mu}, \bar{\mu}) \;\;\leq\;\; (P_{\mathsf{con}} - X^*_{\mathsf{tot}}(\bar{\mu}))(\bar{\mu} - \underline{\mu})\lambda \leq (\bar{\mu} - \underline{\mu})P_{\mathsf{con}}. \quad \text{(B.21)}$$

B.4 Proof of Lemma 4

Let $\tilde{\mu} \in [\mu_{\mathsf{min}}, \mu_{\mathsf{max}}]$ be any value of the Lagrangian dual variable for the CSRA problem. Then, at $\tilde{\mu}$, one of the following three cases hold.

1. For all n, $|S_n(\tilde{\mu})| \leq 1$.
2. For some n, $|S_n(\tilde{\mu})| > 1$ but no two combinations in $S_n(\tilde{\mu})$ have the same allocated power.
3. For some n, $|S_n(\tilde{\mu})| > 1$ and at least two combinations in $S_n(\tilde{\mu})$ have the same allocated power.

We will now study each case individually.

B.4.1 Case I : $|S_n(\tilde{\mu})| \leq 1 \;\forall n$.

In this case, we have $\mathbf{I}^*(\tilde{\mu}) \in \{0,1\}^{N\times K\times M}$ that can be found by using (3.16). Now, for any subchannel n, let us suppose that $|S_n(\tilde{\mu})| = 0$. Then, for all (k,m), we have $V_{n,k,m}(\tilde{\mu}, p^*_{n,k,m}(\tilde{\mu})) > 0$. Since $p^*_{n,k,m}(\mu)$ is a continuous function of μ, $V_{n,k,m}(\mu, p^*_{n,k,m}(\mu))$ is a continuous function of μ. By definition [84], if $f(\cdot)$ is a function continuous at x_0, then for every $\nu > 0$, we can fix a δ such that

$$|f(x) - f(x_0)| < \nu, \text{ whenever } |x - x_0| < \delta.$$

In our setting, therefore, we can fix a $\delta^1_{n,k,m}\,(>0)$ such that

$$\left|V_{n,k,m}(\mu, p^*_{n,k,m}(\mu)) - V_{n,k,m}(\tilde{\mu}, p^*_{n,k,m}(\tilde{\mu}))\right| < \frac{1}{2}\min_{n,k,m} V_{n,k,m}(\tilde{\mu}, p^*_{n,k,m}(\tilde{\mu})),$$

whenever $|\mu - \tilde{\mu}| < \delta^1_{n,k,m}$. This implies

$$V_{n,k,m}(\mu, p^*_{n,k,m}(\mu)) > 0, \text{ whenever } |\mu - \tilde{\mu}| < \delta^1_{n,k,m}. \tag{B.22}$$

In other words, $|S_n(\mu)| = |S_n(\tilde{\mu})| = 0$.

Now, let us suppose that $|S_n(\tilde{\mu})| = 1$ and $(k^*(n), m^*(n)) \in S_n(\tilde{\mu})$. Then, for any other combination (k, m), let us define

$$W_{n,k,m}(\tilde{\mu}) \triangleq V_{n,k,m}(\tilde{\mu}, p^*_{n,k,m}(\tilde{\mu})) - V_{n,k^*(n),m^*(n)}(\tilde{\mu}, p^*_{n,k^*(n),m^*(n)}(\tilde{\mu})).$$

Clearly, $W_{n,k,m}(\mu)$ is a continuous function of μ and takes on a positive value at $\mu = \tilde{\mu}$ for all $(k, m) \neq (k^*(n), m^*(n))$. Therefore, using arguments from the scenario of $|S_n(\tilde{\mu})| = 0$ (see (B.22)), we can fix a $\delta^1_{n,k,m} > 0$ such that

$$W_{n,k,m}(\mu) > 0, \text{ whenever } |\mu - \tilde{\mu}| < \delta^1_{n,k,m},$$

for all $(k, m) \neq (k^*(n), m^*(n))$. In other words, $S_n(\mu) = \{(k^*(n), m^*(n))\}$.

Selecting $\delta_1 = \min_{n,k,m} \delta^1_{n,k,m}$, we have that

$$|S_n(\mu)| \leq 1 \ \forall n, \text{ whenever } |\mu - \tilde{\mu}| < \delta_1. \tag{B.23}$$

Moreover, if $\mu_1, \mu_2 \in (\mu - \delta_1, \mu + \delta_1)$, we have $S_n(\mu_1) = S_n(\mu_2)$ for all n and $|S_n(\mu_1)| \leq 1$. This implies that $\mathbf{I}^*(\mu_1), \mathbf{I}^*(\mu_2) \in \{0, 1\}^{N \times K \times M}$ and $\mathbf{I}^*(\mu_1) = \mathbf{I}^*(\mu_2)$.

B.4.2 Case II : For some n, $|S_n(\tilde{\mu})| > 1$ but no two combinations in $S_n(\tilde{\mu})$ have the same allocated power.

From case I, $\forall (k, m) \notin S_n(\tilde{\mu})$, we can fix a δ_1 (> 0) such that $V_{n,k,m}(\mu, p^*_{n,k,m}(\mu)) > 0$ whenever $|\mu - \tilde{\mu}| < \delta_1$. For other combinations, let us start by giving the following definitions.

$$\begin{aligned} (k^*_{\mathsf{max}}(n), m^*_{\mathsf{max}}(n)) &\triangleq \underset{(k,m) \in S_n(\tilde{\mu})}{\text{argmax}} \ p^*_{n,k,m}(\tilde{\mu}), \text{ and} \\ (k^*_{\mathsf{min}}(n), m^*_{\mathsf{min}}(n)) &\triangleq \underset{(k,m) \in S_n(\tilde{\mu})}{\text{argmin}} \ p^*_{n,k,m}(\tilde{\mu}). \end{aligned}$$

We will now show that we can fix a δ_2 (> 0) such that $S_n(\mu) = \{(k^*_{\mathsf{max}}(n), m^*_{\mathsf{max}}(n))\}$ whenever $\mu \in (\tilde{\mu}-\delta_2, \tilde{\mu})$ and whenever $\mu \in (\tilde{\mu}, \tilde{\mu}+\delta_2)$, either $S_n(\mu) = \{(k^*_{\mathsf{min}}(n), m^*_{\mathsf{min}}(n))\}$ or $S_n(\mu) = \phi$.

Consider the case when $\mu < \tilde{\mu}$. Let us define

$$T_{n,k,m}(\mu, \tilde{\mu}) \triangleq \frac{V_{n,k,m}(\mu, p^*_{n,k,m}(\mu)) - V_{n,k,m}(\tilde{\mu}, p^*_{n,k,m}(\tilde{\mu}))}{\mu - \tilde{\mu}}.$$

Note that $\lim_{\mu\to\tilde{\mu}} T_{n,k,m}(\mu, \tilde{\mu}) = \frac{\partial V_{n,k,m}(\mu, p^*_{n,k,m}(\mu))}{\partial \mu}\Big|_{\mu=\tilde{\mu}}$, where

$$\begin{aligned}
&\frac{\partial V_{n,k,m}(\mu, p^*_{n,k,m}(\mu))}{\partial \mu} \\
&= p^*_{n,k,m}(\mu) + \mu\frac{\partial p^*_{n,k,m}(\mu)}{\partial \mu} \qquad \text{(B.24)} \\
&\quad -a_m b_m r_m \,\mathrm{E}\left\{U'_{n,k,m}\big((1 - a_m e^{-b_m p^*_{n,k,m}(\mu)\gamma_{n,k}})r_m\big)\gamma_{n,k} e^{-b_m p^*_{n,k,m}(\mu)\gamma_{n,k}}\right\}\frac{\partial p^*_{n,k,m}(\mu)}{\partial \mu}.
\end{aligned}$$

Now, if (3.11) is satisfied for a non-negative value of $\tilde{p}_{n,k,m}(\mu)$, i.e., $p^*_{n,k,m}(\mu) = \tilde{p}_{n,k,m}(\mu)$, then

$$\lim_{\mu\to\tilde{\mu}} T_{n,k,m}(\mu, \tilde{\mu}) = \frac{\partial V_{n,k,m}(\mu, p^*_{n,k,m}(\mu))}{\partial \mu} = p^*_{n,k,m}(\mu). \qquad \text{(B.25)}$$

On the contrary, if (3.11) is not satisfied for a non-negative value of $\tilde{p}_{n,k,m}(\mu)$, then $\mu > a_m b_m r_m U'_{n,k,m}\big((1-a_m)r_m\big)\,\mathrm{E}\{\gamma_{n,k}\}$. Since the RHS of (3.11) is a continuous function of $\tilde{p}_{n,k,m}(\mu)$, $p^*_{n,k,m}(\mu) = 0$ in a small neighborhood around μ giving $\frac{\partial p^*_{n,k,m}(\mu)}{\partial \mu} = 0$. Therefore, (B.25) is true for all possible values of μ.

Then, for any $(k, m) \in S_n(\tilde{\mu}) \setminus \{(k^*_{\mathsf{max}}(n), m^*_{\mathsf{max}}(n))\}$, we have

$$\begin{aligned}
&\lim_{\mu\to\tilde{\mu}} \frac{V_{n,k^*_{\mathsf{max}}(n),m^*_{\mathsf{max}}(n)}(\mu, p^*_{n,k^*_{\mathsf{max}}(n),m^*_{\mathsf{max}}(n)}(\mu)) - V_{n,k,m}(\mu, p^*_{n,k,m}(\mu))}{\mu - \tilde{\mu}} \qquad \text{(B.26)} \\
&\qquad = \lim_{\mu\to\tilde{\mu}} T_{n,k^*_{\mathsf{max}}(n),m^*_{\mathsf{max}}(n)}(\mu, \tilde{\mu}) - T_{n,k,m}(\mu, \tilde{\mu}) \\
&\qquad = p^*_{n,k^*_{\mathsf{max}}(n),m^*_{\mathsf{max}}(n)}(\tilde{\mu}) - p^*_{n,k,m}(\tilde{\mu}) \qquad \text{(B.27)} \\
&\qquad > 0,
\end{aligned}$$

where (B.27) follows from (B.25). Note that the denominator in (B.26) is less than zero. Therefore, we can fix a $\delta^1_{n,k,m}$ (> 0) such that

$$V_{n,k^*_{\max}(n),m^*_{\max}(n)}(\mu, p_{n,k^*_{\max}(n),m^*_{\max}(n)}(\mu)) < V_{n,k,m}(\mu, p^*_{n,k,m}(\mu)), \tag{B.28}$$

whenever $0 < \tilde{\mu} - \mu < \delta^1_{n,k,m}$. Since, this is true for all $(k,m) \in S_n(\tilde{\mu}) \backslash \{(k^*_{\max}(n), m^*_{\max}(n))\}$, selecting $\delta^1 = \min\{\delta_1, \min_{n,k,m} \delta^1_{n,k,m}\}$, we have $S_n(\mu) = \{(k^*_{\max}(n), m^*_{\max}(n))\}$ whenever $\mu \in (\tilde{\mu} - \delta^1, \tilde{\mu})$. Note that $V_{n,k,m}(\mu, p^*_{n,k,m}(\mu))$ being an increasing function of μ, implying that

$$V_{n,k^*_{\max}(n),m^*_{\max}(n)}(\mu, p_{n,k^*_{\max}(n),m^*_{\max}(n)}(\mu)) \leq V_{n,k^*_{\max}(n),m^*_{\max}(n)}(\tilde{\mu}, p_{n,k^*_{\max}(n),m^*_{\max}(n)}(\tilde{\mu})) \leq 0.$$

Now, consider that $\mu > \tilde{\mu}$. Using the arguments from the above discussion, we obtain that for any $(k,m) \in S_n(\tilde{\mu}) \setminus \{(k^*_{\min}(n), m^*_{\min}(n))\}$, there exists a $\delta^2_{n,k,m}$, such that

$$V_{n,k^*_{\min}(n),m^*_{\min}(n)}(\mu, p_{n,k^*_{\min}(n),m^*_{\min}(n)}(\mu)) < V_{n,k,m}(\mu, p^*_{n,k,m}(\mu)),$$

whenever $0 < \mu - \tilde{\mu} < \delta^2_{n,k,m}$. Let $V_{n,k,m}(\tilde{\mu}, p^*_{n,k,m}(\tilde{\mu})) < 0$. Then we can fix a δ_3 (> 0) such that

$$V_{n,k^*_{\min}(n),m^*_{\min}(n)}(\mu, p_{n,k^*_{\min}(n),m^*_{\min}(n)}(\mu)) < 0,$$

whenever $0 \leq \mu - \tilde{\mu} < \delta_3$. Selecting $\delta^2 = \min\{\delta_1, \min_{n,k,m} \delta^2_{n,k,m}, \delta_3\}$, we then have $S_n(\mu) = \{(k^*_{\min}(n), m^*_{\min}(n))\}$ whenever $\mu \in (\tilde{\mu}, \tilde{\mu} + \delta^2)$. However, if $V_{n,k,m}(\tilde{\mu}, p^*_{n,k,m}(\tilde{\mu})) = 0$, then selecting $\delta^2 = \min\{\delta_1, \min_{n,k,m} \delta^2_{n,k,m}\}$, we have $S_n(\mu) = \phi$ whenever $\mu \in (\tilde{\mu}, \tilde{\mu} + \delta^2)$. In both situations, if $\mu_1, \mu_2 \in (\tilde{\mu}, \tilde{\mu} + \delta^2)$, we get $S_n(\mu_1) = S_n(\mu_2)$.

Summarizing the discussion for case II, selecting $\delta_2 = \min\{\delta^1, \delta^2\}$, we have $\mathbf{I}^*(\mu) \in \{0,1\}^{N \times K \times M}$ whenever $|\mu - \tilde{\mu}| < \delta^2,\ \mu \neq \tilde{\mu}$. Furthermore, if $\mu_1, \mu_2 \in (\tilde{\mu} - \delta_2, \tilde{\mu})$ or if $\mu_1, \mu_2 \in (\tilde{\mu}, \tilde{\mu} + \delta_2)$, then $S_n(\mu_1) = S_n(\mu_2)\ \forall n$ and $\mathbf{I}^*(\mu_1) = \mathbf{I}^*(\mu_2)$.

B.4.3 Case III : For some n, $|S_n(\tilde{\mu})| > 1$ and at least two combinations in $S_n(\tilde{\mu})$ have the same allocated power.

Let us say that (k_1, m_1) and (k_2, m_2) lie in the set $S_n(\tilde{\mu})$ for some n and $p^*_{n,k_1,m_1}(\tilde{\mu}) = p^*_{n,k_2,m_2}(\tilde{\mu})$. Let us also assume that $I^*_{n,k_1,m_1}(\tilde{\mu})$ and $I^*_{n,k_2,m_2}(\tilde{\mu})$ are the corresponding optimal allocations. Then, their contribution towards the optimal $L_n(\tilde{\mu}, \cdot)$ is

$$\begin{aligned} & I^*_{n,k_1,m_1}(\tilde{\mu})V_{n,k_1,m_1}(\tilde{\mu}, p^*_{n,k_1,m_1}(\tilde{\mu})) + I^*_{n,k_2,m_2}(\tilde{\mu})V_{n,k_2,m_2}(\tilde{\mu}, p^*_{n,k_2,m_2}(\tilde{\mu})) \\ = \; & (I^*_{n,k_1,m_1}(\tilde{\mu}) + I^*_{n,k_2,m_2}(\tilde{\mu}))V_{n,k_1,m_1}(\tilde{\mu}, p^*_{n,k_1,m_1}(\tilde{\mu})). \end{aligned}$$

Moreover, the total power allocated is

$$\begin{aligned} & I^*_{n,k_1,m_1}(\tilde{\mu})p^*_{n,k_1,m_1}(\tilde{\mu}) + I^*_{n,k_2,m_2}(\tilde{\mu})p^*_{n,k_2,m_2}(\tilde{\mu}) \\ = \; & (I^*_{n,k_1,m_1}(\tilde{\mu}) + I^*_{n,k_2,m_2}(\tilde{\mu}))p^*_{n,k_1,m_1}(\tilde{\mu}). \end{aligned} \tag{B.29}$$

Therefore, we can ignore (k_2, m_2) and give its allocation to (k_1, m_1) while maintaining optimality. Mathematically, the allocation $\tilde{I}^*(\tilde{\mu})$ defined by

$$\tilde{I}^*_{n,k,m}(\tilde{\mu}) = \begin{cases} I^*_{n,k,m}(\tilde{\mu}) & (n,k,m) \notin \{(n,k_1,m_1),(n,k_2,m_2)\} \\ I^*_{n,k_1,m_1}(\tilde{\mu}) + I^*_{n,k_2,m_2}(\tilde{\mu}) & (n,k,m) = (n,k_1,m_1) \\ 0 & (n,k,m) = (n,k_2,m_2), \end{cases} \tag{B.30}$$

also achieves the optimal Lagrangian value at $\mu = \tilde{\mu}$ and has the same allocated sum-power. Note that this can be repeated for every subchannel n and for each n, over all such combinations with equal allocated powers.

Combining cases I, II, and III, we can fix a $\delta = \min\{\delta_1, \delta_2\}$ for every $\tilde{\mu}$ such that $\mathbf{I}^*(\mu) \in \{0,1\}^{N\times K\times M}$ whenever $|\mu - \tilde{\mu}| < \delta$, $\mu \neq \tilde{\mu}$. Moreover, for all $\mu_1, \mu_2 \in (\tilde{\mu} - \delta, \tilde{\mu})$, there exists $\mathbf{I}^*(\mu_1), \mathbf{I}^*(\mu_2) \in \{0,1\}^{N\times K\times M}$ such that $\mathbf{I}^*(\mu_1) = \mathbf{I}^*(\mu_2)$. The same property holds when both $\mu_1, \mu_2 \in (\tilde{\mu}, \tilde{\mu} + \delta)$.

B.5 Proof of Lemma 5

The proof involves the use of generalized Lagrange multiplier method. From (3.6), we have

$$\Big(\mathbf{I}^*(\mu), \mathbf{x}^*(\mu, \mathbf{I}^*(\mu))\Big) = \underset{\substack{\mathbf{x}\succeq 0 \\ \mathbf{I}\in\mathcal{I}_{\mathsf{CSRA}}}}{\operatorname{argmin}} \sum_{n,k,m} I_{n,k,m}\, F_{n,k,m}(I_{n,k,m}, x_{n,k,m}) + \Big(\sum_{n,k,m} x_{n,k,m} - P_{\mathsf{con}}\Big)\mu, \tag{B.31}$$

where $F_{n,k,m}(\cdot,\cdot)$ is defined in (3.4). Since, $\mathbf{I}^*(\mu) \in \{0,1\}^{N\times K\times M}$ by assumption and $\mathbf{I}^*(\mu) \in \mathcal{I}_{\mathsf{CSRA}}$, we have $\mathbf{I}^*(\mu) \in \mathcal{I}_{\mathsf{DSRA}} \subset \mathcal{I}_{\mathsf{CSRA}}$. Therefore,

$$\Big(\mathbf{I}^*(\mu), \mathbf{x}^*(\mu, \mathbf{I}^*(\mu))\Big) = \underset{\substack{\mathbf{x}\succeq 0 \\ \mathbf{I}\in\mathcal{I}_{\mathsf{DSRA}}}}{\operatorname{argmin}} \sum_{n,k,m} I_{n,k,m}\, F_{n,k,m}(I_{n,k,m}, x_{n,k,m}) + \Big(\sum_{n,k,m} x_{n,k,m} - P_{\mathsf{con}}\Big)\mu. \tag{B.32}$$

Reference [85] clearly states that the theory of Lagrangian extends to arbitrary real-valued functions (differentiable or not) over arbitrary constraint sets. In particular, applying [85, Theorem 1] to (B.32), we deduce that $\mathbb{I}^* = \mathbf{I}^*(\mu)$ and $\mathbb{X}^* = \mathbf{x}^*(\mu, \mathbf{I}^*(\mu)$, where $(\mathbb{I}^*, \mathbb{X}^*)$ solves the following (primal) problem

$$\begin{aligned}(\mathbb{I}^*, \mathbb{X}^*) \;=\; & \underset{\substack{\{\mathbb{X}\succeq 0\} \\ \mathbb{I}\in\mathcal{I}_{\mathsf{DSRA}}}}{\operatorname{argmin}} \sum_{n,k,m} \mathbb{I}_{n,k,m} F_{n,k,m}(\mathbb{I}_{n,k,m}, \mathbb{X}_{n,k,m}) \\ & \text{s.t.} \quad \sum_{n,k,m} \mathbb{X}_{n,k,m} \le \sum_{n,k,m} x^*_{n,k,m}(\mu, \mathbf{I}^*(\mu)).\end{aligned} \tag{B.33}$$

Substituting back $\mathbb{X}_{n,k,m} = \mathbb{I}_{n,k,m}\mathbb{P}_{n,k,m}$ in the above equation, we get

$$\begin{aligned}(\mathbf{I}^*(\mu), \mathbf{p}^*(\mu, \mathbf{I}^*(\mu)) \;=\; & \underset{\substack{\{\mathbb{X}\succeq 0\} \\ \mathbb{I}\in\mathcal{I}_{\mathsf{DSRA}}}}{\operatorname{argmin}} - \sum_{n,k,m} \mathbb{I}_{n,k,m}\, \mathrm{E}\left\{U_{n,k,m}\big((1 - a_m e^{-b_m \mathbb{P}_{n,k,m}\gamma_{n,k}})r_m\big)\right\} \\ & \text{s.t.} \quad \sum_{n,k,m} \mathbb{I}_{n,k,m}\mathbb{P}_{n,k,m} \le \sum_{n,k,m} x^*_{n,k,m}(\mu, \mathbf{I}^*(\mu)),\end{aligned} \tag{B.34}$$

where

$$p^*_{n,k,m}(\mu, \mathbf{I}^*(\mu)) = \begin{cases} \frac{x^*_{n,k,m}(\mu, \mathbf{I}^*(\mu))}{I^*_{n,k,m}(\mu)} & \text{if } I^*_{n,k,m}(\mu) \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{B.35}$$

B.6 Proof of Lemma 6

Let us denote $\lim_{\underline{\mu}\to\bar{\mu}} \hat{U}_{\mathsf{DSRA}}(\underline{\mu},\bar{\mu})$ by $\hat{U}_{\mathsf{DSRA}}$. The left inequality in the lemma is straightforward since $U^*_{\mathsf{DSRA}} \geq \hat{U}_{\mathsf{DSRA}}(\underline{\mu},\bar{\mu})\ \forall \underline{\mu},\bar{\mu}$. Now, if $|S_n(\mu^*)| \leq 1\ \forall n$, then we have $U^*_{\mathsf{DSRA}} = U^*_{\mathsf{CSRA}} = \hat{U}_{\mathsf{DSRA}}$, ensuring that the solution obtained via the proposed DSRA algorithm is optimal in the limit $\underline{\mu},\bar{\mu} \to \mu^*$. However, when $|S_n(\mu^*)| > 1$ for some n, P_{con} lies in one of the "gaps" as mentioned in Fig. 3.3 and $\mathbf{I}^*_{\mathsf{CSRA}} \notin \mathcal{I}_{\mathsf{DSRA}}$. In this case, we have $0 \leq U^*_{\mathsf{DSRA}} - \hat{U}_{\mathsf{DSRA}} \leq U^*_{\mathsf{CSRA}} - \hat{U}_{\mathsf{DSRA}}$. Let $U^*(\mathbf{I})$ be the optimal utility achieved for user-MCS allocation matrix $\mathbf{I} \in \mathcal{I}_{\mathsf{DSRA}}$. We recall from Section 3.4.3 that, at μ^*, the allocation $\mathbf{I}^{\mathsf{min}}(\mu^*)$ is one of possibly many values of $\mathbf{I}$ minimizing $L(\mu^*,\mathbf{I},\mathbf{x}^*(\mu^*,\mathbf{I}))$. Thus, $U^*_{\mathsf{CSRA}} = -L(\mu^*,\mathbf{I}^{\mathsf{min}}(\mu^*),\mathbf{x}^*(\mu^*,\mathbf{I}^{\mathsf{min}}(\mu^*)))$. For brevity in this proof, let us denote $\mathbf{I}^{\mathsf{min}}(\mu^*)$ and $\mathbf{I}^{\mathsf{max}}(\mu^*)$ ($\in \mathcal{I}_{\mathsf{DSRA}}$), defined in (3.30), by $\mathbf{I}^{\mathsf{min}}$ and $\mathbf{I}^{\mathsf{max}}$, respectively. Therefore, $\hat{U}_{\mathsf{DSRA}} = \max\{U^*(\mathbf{I}^{\mathsf{min}}), U^*(\mathbf{I}^{\mathsf{max}})\}$. This gives us

$$\begin{aligned} U^*_{\mathsf{CSRA}} - \hat{U}_{\mathsf{DSRA}} &\leq U^*_{\mathsf{CSRA}} - U^*(\mathbf{I}^{\mathsf{min}}) \\ &= -L(\mu^*,\mathbf{I}^{\mathsf{min}},\mathbf{x}^*(\mu^*,\mathbf{I}^{\mathsf{min}})) + L_{\mathbf{I}^{\mathsf{min}}}(\mu^*_{\mathbf{I}^{\mathsf{min}}},\mathbf{x}^*(\mu^*_{\mathbf{I}^{\mathsf{min}}}))) \\ &= -L(\mu^*,\mathbf{I}^{\mathsf{min}},\mathbf{x}^*(\mu^*,\mathbf{I}^{\mathsf{min}})) + L(\mu^*_{\mathbf{I}^{\mathsf{min}}},\mathbf{I}^{\mathsf{min}},\mathbf{x}^*(\mu^*_{\mathbf{I}^{\mathsf{min}}},\mathbf{I}^{\mathsf{min}})), \quad \text{(B.36)} \end{aligned}$$

where, for (B.36), we use the equivalence between $L(\mu,\mathbf{I},\mathbf{x})$ in (3.5) and $L_{\mathbf{I}}(\mu,\mathbf{x})$ in (3.36). Note that $\mu^*_{\mathbf{I}^{\mathsf{min}}} \leq \mu^*$, since the total optimally allocated power for $\mathbf{I}^{\mathsf{min}}$ at $\mu = \mu^*$ is less than or equal to P_{con} and the total optimally allocated power for any

given $\mathbf{I}$ is a decreasing function of μ. Plugging $L(\cdot,\cdot,\cdot)$ from (3.5) into (B.36), we get

$$\begin{aligned} & U^*_{\mathsf{CSRA}} - \hat{U}_{\mathsf{DSRA}} \\ & \leq \; -\Big[-\mu^* P_{\mathsf{con}} + \sum_{n,k,m} I^{\mathsf{min}}_{n,k,m} \Big(-\bar{U}_{n,k,m}(p^*_{n,k,m}(\mu^*)) + \mu^* p^*_{n,k,m}(\mu^*) \Big) \Big] \qquad \text{(B.37)} \\ & \quad + \Big[-\mu^*_{\mathbf{I}^{\mathsf{min}}} P_{\mathsf{con}} + \sum_{n,k,m} I^{\mathsf{min}}_{n,k,m} \Big(-\bar{U}_{n,k,m}(p^*_{n,k,m}(\mu^*_{\mathbf{I}^{\mathsf{min}}})) + \mu^*(\mathbf{I}^{\mathsf{min}}) p^*_{n,k,m}(\mu^*_{\mathbf{I}^{\mathsf{min}}}) \Big) \Big], \end{aligned}$$

where, $\bar{U}_{n,k,m}(x) = \mathrm{E}\left\{ U_{n,k,m}\big((1 - a_{k,m} e^{-b_{k,m} x \gamma_{n,k}}) r_{k,m}\big) \right\}$. Using the definition of $X^*_{\mathsf{tot}}(\mathbf{I}, \mu)$ in (3.41), we have $X^*_{\mathsf{tot}}(\mathbf{I}^{\mathsf{min}}, \mu^*) \leq P_{\mathsf{con}}$ and $X^*_{\mathsf{tot}}(\mathbf{I}^{\mathsf{min}}, \mu^*_{\mathbf{I}^{\mathsf{min}}}) = P_{\mathsf{con}}$. Therefore, (B.37) can be re-written as

$$\begin{aligned} & U^*_{\mathsf{CSRA}} - \hat{U}_{\mathsf{DSRA}} \qquad \text{(B.38)} \\ & \leq \; \mu^* \big(P_{\mathsf{con}} - X^*_{\mathsf{tot}}(\mathbf{I}^{\mathsf{min}}, \mu^*) \big) - \sum_{n,k,m} I^{\mathsf{min}}_{n,k,m} \Big[\bar{U}_{n,k,m}(p^*_{n,k,m}(\mu^*_{\mathbf{I}^{\mathsf{min}}})) - \bar{U}_{n,k,m}(p^*_{n,k,m}(\mu^*)) \Big]. \end{aligned}$$

Calculating the first two derivatives of $\bar{U}_{n,k,m}(x)$ with respect to x, we find that it is a strictly-increasing concave function of x. Therefore, if $x_1 \leq x_2$, one can write that $\bar{U}_{n,k,m}(x_2) - \bar{U}_{n,k,m}(x_1) \geq (x_2 - x_1) \bar{U}'_{n,k,m}(x_2)$. Plugging $x_1 = p^*_{n,k,m}(\mu^*)$ and $x_2 = p^*_{n,k,m}(\mu^*_{\mathbf{I}^{\mathsf{min}}})$ into this inequality, we get

$$\begin{aligned} & \bar{U}_{n,k,m}(p^*_{n,k,m}(\mu^*_{\mathbf{I}^{\mathsf{min}}})) - \bar{U}_{n,k,m}(p^*_{n,k,m}(\mu^*)) \\ & \quad \geq \; \Big(p^*_{n,k,m}(\mu^*_{\mathbf{I}^{\mathsf{min}}}) - p^*_{n,k,m}(\mu^*) \Big) \frac{\partial \bar{U}_{n,k,m}(x)}{\partial x} \Big|_{x = p^*_{n,k,m}(\mu^*_{\mathbf{I}^{\mathsf{min}}})} \qquad \text{(B.39)} \end{aligned}$$

From (B.38) and (B.39), we then get

$$\begin{aligned} & U^*_{\mathsf{CSRA}} - \hat{U}_{\mathsf{DSRA}} \qquad \text{(B.40)} \\ & \leq \; \mu^* \big(P_{\mathsf{con}} - X^*_{\mathsf{tot}}(\mathbf{I}^{\mathsf{min}}, \mu^*) \big) - \sum_{n,k,m} I^{\mathsf{min}}_{n,k,m} \bar{U}'_{n,k,m}\big(p^*_{n,k,m}(\mu^*_{\mathbf{I}^{\mathsf{min}}})\big) \Big(p^*_{n,k,m}(\mu^*_{\mathbf{I}^{\mathsf{min}}}) - p^*_{n,k,m}(\mu^*) \Big). \end{aligned}$$

Evaluating $\bar{U}'_{n,k,m}\big(p^*_{n,k,m}(\mu^*_{\mathbf{I}^{\mathsf{min}}})\big)$, we find

$$\begin{aligned}\left.\frac{\partial \bar{U}_{n,k,m}(x)}{\partial x}\right|_{x=p^*_{n,k,m}(\mu^*_{\mathbf{I}^{\mathsf{min}}})} &= a_{k,m}b_{k,m}r_{k,m}\,\mathrm{E}\left\{U'_{n,k,m}\big((1-a_{k,m}e^{-b_{k,m}p^*_{n,k,m}(\mu^*_{\mathbf{I}^{\mathsf{min}}})\gamma_{n,k}})r_{k,m}\big)\right. \\ &\qquad \left.\times\,\gamma_{n,k}e^{-b_{k,m}p^*_{n,k,m}(\mu^*_{\mathbf{I}^{\mathsf{min}}})\gamma_{n,k}}\right\} \\ &\geq \mu_{\mathsf{min}}. \end{aligned} \tag{B.41}$$

From (B.40) and (B.41), we finally obtain

$$\begin{aligned} U^*_{\mathsf{CSRA}} - \hat{U}_{\mathsf{DSRA}} &\leq (\mu^* - \mu_{\mathsf{min}})\big(P_{\mathsf{con}} - X^*_{\mathsf{tot}}(\mathbf{I}^{\mathsf{min}}, \mu^*)\big) \\ &\leq (\mu_{\mathsf{max}} - \mu_{\mathsf{min}})P_{\mathsf{con}}. \end{aligned} \tag{B.42}$$

Appendix C: Proofs in Chapter 5

C.1 Proof of Theorem 1

By ignoring the interference, we have

$$\mathcal{C}_{\mathbf{x},\mathbf{y},\nu}(\mathbf{U},\mathbf{P}) \;\leq\; \sum_{i=1}^{B}\sum_{n=1}^{N}\log\Big(1+P_{i,n}\,\gamma_{i,U_{i,n},n}\Big) \tag{C.1}$$

Taking expectation w.r.t. $\{\mathbf{x},\mathbf{y},\nu\}$, we have

$$\begin{aligned}
\mathcal{C}^* &= \mathrm{E}\{\mathcal{C}_{\mathbf{x},\mathbf{y},\nu}(\mathbf{U},\mathbf{P})\} \\
&\leq \sum_{i=1}^{B}\sum_{n=1}^{N}\max_k \mathrm{E}\Big\{\log\big(1+P_{\mathsf{con}}\,\gamma_{i,k,n}\big)\Big\} \\
&\leq \sum_{i=1}^{B}\sum_{n=1}^{N}\mathrm{E}\Big\{\max_k \log\big(1+P_{\mathsf{con}}\gamma_{i,k,n}\big)\Big\} &\text{(C.2)}\\
&\leq \sum_{i=1}^{B}\sum_{n=1}^{N}\mathrm{E}\Big\{\log\big(1+P_{\mathsf{con}}\max_k \gamma_{i,k,n}\big)\Big\}, &\text{(C.3)}
\end{aligned}$$

where (C.2) follows because, for any function $f(\cdot,\cdot)$, $\max_k \mathrm{E}\{f(k,\cdot)\} \leq \mathrm{E}\{\max_k f(k,\cdot)\}$, and (C.3) follows because $\log(\cdot)$ is a non-decreasing function. One can also construct an alternate upper bound by applying Jensen's inequality to the RHS of (C.1) as follows:

$$\begin{aligned}
\mathcal{C}_{\mathbf{x},\mathbf{y},\nu}(\mathbf{U},\mathbf{P}) &\leq N\sum_{i=1}^{B}\log\Big(1+\frac{1}{N}\sum_n P_{i,n}\,\gamma_{i,U_{i,n},n}\Big) &\text{(C.4)}\\
&\leq N\sum_{i=1}^{B}\log\Big(1+\frac{P_{\mathsf{con}}}{N}\max_{n,k}\gamma_{i,k,n}\Big), &\text{(C.5)}
\end{aligned}$$

since $\sum_n P_{i,n} \leq P_{\mathsf{con}}$. Therefore,

$$\begin{aligned} \mathcal{C}^* &= \mathrm{E}\{\mathcal{C}_{\mathbf{x},\mathbf{y},\nu}(\mathbf{U},\mathbf{P})\} \\ &\leq N \sum_{i=1}^{B} \mathrm{E}\left\{ \log\left(1 + \frac{P_{\mathsf{con}}}{N} \max_{n,k} \gamma_{i,k,n}\right)\right\}. \end{aligned} \tag{C.6}$$

Combining (C.3) and (C.6), we obtain

$$\begin{aligned} \mathcal{C}^* \leq \min\Bigg\{ & \sum_{i,n} \mathrm{E}_{\mathbf{x},\mathbf{y},\nu}\left\{ \log\left(1 + P_{\mathsf{con}} \max_{k} \gamma_{i,k,n}\right)\right\}, \\ & N \sum_{i} \mathrm{E}_{\mathbf{x},\mathbf{y},\nu}\left\{ \log\left(1 + \frac{P_{\mathsf{con}}}{N} \max_{n,k} \gamma_{i,k,n}\right)\right\}\Bigg\}. \end{aligned} \tag{C.7}$$

For lower bound, let P_{con}/N power be allocated to each resource-block by every BS. Then,

$$\mathcal{C}_{\mathbf{x},\mathbf{y},\nu}(\mathbf{U},\mathbf{P}) \geq \sum_{i=1}^{B}\sum_{n=1}^{N} \log\left(1 + \frac{P_{\mathsf{con}}\,\gamma_{i,k_{i,n},n}}{N + P_{\mathsf{con}} \sum_{j\neq i} \gamma_{j,k_{i,n},n}}\right), \tag{C.8}$$

where $k_{i,n}$ is any other user allocated on subchannel n by BS i. Note that, due to sub-optimal power allocation, all user-allocation strategies $\{k_{i,n}, \forall i, n\}$ achieve a utility that is lower that $\mathcal{C}_{\mathbf{x},\mathbf{y},\nu}(\mathbf{U},\mathbf{P})$. To handle (C.8) easily, we introduce an indicator variable $I_{i,k,n}(\mathbf{x},\mathbf{y},\nu)$ which equals 1 if $k = k_{i,n}$, otherwise takes the value 0. Since, each BS i can schedule at-most one user on any resource block n in a given time-slot, we have $\sum_k I_{i,k,n}(\mathbf{x},\mathbf{y},\nu) = 1 \;\forall\; i, n$. Now, (C.8) can be re-written as:

$$\mathcal{C}_{\mathbf{x},\mathbf{y},\nu}(\mathbf{U},\mathbf{P}) \geq \sum_{i=1}^{B}\sum_{n=1}^{N}\sum_{k=1}^{K} I_{i,k,n}(\mathbf{x},\mathbf{y},\nu) \log\left(1 + \frac{P_{\mathsf{con}}\,\gamma_{i,k,n}}{N + P_{\mathsf{con}} \sum_{j\neq i} \gamma_{j,k,n}}\right).$$

Taking expectation w.r.t. $(\mathbf{x},\mathbf{y},\nu)$, we get

$$\begin{aligned} \mathcal{C}^* &\geq \sum_{i,n,k} \mathrm{E}\left\{ I_{i,k,n}(\mathbf{x},\mathbf{y},\mathbf{z}) \log\left(1 + \frac{P_{\mathsf{con}}\,\gamma_{i,k,n}}{N + P_{\mathsf{con}} \sum_{j\neq i} \gamma_{j,k,n}}\right)\right\} \\ &\geq \sum_{i,n,k} \mathrm{E}\left\{ I_{i,k,n}(\mathbf{x},\mathbf{y},\mathbf{z}) \frac{\log\left(1 + P_{\mathsf{con}}\,\gamma_{i,k,n}\right)}{N + P_{\mathsf{con}} \sum_{j\neq i} \gamma_{j,k,n}}\right\}. \end{aligned} \tag{C.9}$$

Here, the last equation holds because for any non-decreasing concave function $V(\cdot)$ (for example, $V(x) = \log(1+x)$) and for all $d_1, d_2 > 0$, we have

$$\begin{aligned} V(d_1) - V(0) &\leq \Big[V\Big(\frac{d_1}{d_2}\Big) - V(0)\Big] d_2 \\ \implies V\Big(\frac{d_1}{d_2}\Big) &\geq \frac{V(d_1) - V(0)}{d_2} + V(0). \end{aligned} \tag{C.10}$$

Now, $\frac{1}{N + P_{\mathsf{con}} \sum_{j \neq i} \gamma_{j,k,n}\left(y_{j,k,n}\right)} \leq 1$. Therefore,

$$\mathcal{C}^* \geq \sum_{i,n,k} \mathrm{E}\left\{ \frac{I_{i,k,n}(\mathbf{x}, \mathbf{y}, \nu) \log\left(1 + P_{\mathsf{con}}\, \gamma_{i,k,n}\right)}{N + P_{\mathsf{con}} \sum_{j \neq i} \gamma_{j,k,n}} \right\}. \tag{C.11}$$

To obtain the best lower bound, we now select the user $k_{i,n}$ to be the one for which $\gamma_{i,k,n}$ attains the highest value for every combination (i,n), i.e.,

$$I_{i,k,n}(\mathbf{x}, \mathbf{y}, \nu) = \begin{cases} 1 & \text{if } k = \arg\max_{k'} \gamma_{i,k',n} \\ 0 & \text{otherwise.} \end{cases} \tag{C.12}$$

Using (C.12) in (C.11), we get the lower bound in Theorem 1.

C.2 Proof of Theorem 2 and Theorem 3

The proof outline is as follows. We first prove three lemmas. The first lemma, i.e, Lemma 11, uses one-sided variant of Chebyshev's inequality (also called Cantelli's inequality) and Theorem 1 to show that

$$\mathcal{C}^* \geq f_{\mathsf{lo}}^{\mathsf{DN}}(r, B, N) \sum_{i,n} \mathrm{E}\Big\{ \log\left(1 + P_{\mathsf{con}} \max_k \gamma_{i,k,n}\right) \Big\},$$

where $\mathcal{C}^*$ is expected achievable sum-rate of the system. The second lemma, i.e, Lemma 12, finds the cumulative distribution function (CDF) of channel-SNR, denoted by $F_{\gamma_{i,k,n}}(\cdot)$, under Rayleigh-distributed $|\nu_{i,k,n}|$ and a truncated path-loss model. The third lemma, i.e, Lemma 13, uses Lemma 12 and extreme-value theory to show

that $(\max_k \gamma_{i,k,n} - l_K)$ converges in distribution to a limiting random variable with a Gumbel type cdf, that is given by

$$\exp(-e^{-xr_0^{2\alpha}/\beta^2}),\ x \in (-\infty, \infty), \tag{C.13}$$

where $F_{\gamma_{i,k,n}}(l_K) = 1 - \frac{1}{K}$. Thereafter, we use Theorem 1, Lemma 11, Lemma 13, and [67, Theorem A.2] to obtain the final result.

Now, we give details of the full proof.

Lemma 11. *In a dense-network, the expected achievable sum-rate is lower bounded as:*

$$\mathcal{C}^* \ \geq \ f_{\mathsf{lo}}^{\mathsf{DN}}(r, B, N) \sum_{i,n} \mathrm{E}\Big\{ \log\big(1 + P_{\mathsf{con}} \max_k \gamma_{i,k,n}\big)\Big\}, \tag{C.14}$$

where $r > 0$ is a fixed number, $f_{lo}^{DN}(r, B, N) = \frac{r^2}{(1+r^2)(N+P_{\mathsf{con}}\beta^2 r_0^{-2\alpha}(\mu+r\sigma)B)}$, μ and σ are the mean and standard-deviation of $|\nu_{i,k,n}|^2$.

Proof. We know that

$$\sum_{j\neq i} \gamma_{j,k,n} = \beta^2 \sum_{j\neq i} R_{j,k}^{-2\alpha} |\nu_{j,k,n}|^2 \leq \beta^2 r_0^{-2\alpha} \sum_{j\neq i} |\nu_{j,k,n}|^2. \tag{C.15}$$

Therefore, the lower bound in Theorem 1 reduces to the following equation.

$$\mathcal{C}^* \geq \sum_{i,n,k} \mathrm{E}\left\{ \frac{\max_k \log\big(1 + P_{\mathsf{con}}\gamma_{i,k,n}\big)}{N + P_{\mathsf{con}}\beta^2 r_0^{-2\alpha} \sum_{j\neq i} |\nu_{j,k,n}|^2} \right\}. \tag{C.16}$$

Now, we apply one-sided variant of Chebyshev's inequality (also called Cantelli's inequality) to the term $\sum_{j\neq i} |\nu_{j,k,n}|^2$ in the denominator. By assumption, $|\nu_{i,k,n}|^2$ are i.i.d. across i, k, n with mean μ and variance σ. Hence, applying Cantelli's inequality, We have

$$\begin{aligned}
\Pr\Big(\sum_{j\neq i} |\nu_{j,k,n}|^2 > (B-1)(\mu + r\sigma)\Big) \ &\leq \ \frac{1}{1+r^2} \\
\implies \Pr\Big(\sum_{j\neq i} |\nu_{j,k,n}|^2 > (\mu + r\sigma)B\Big) \ &\leq \ \frac{1}{1+r^2} \qquad \text{(C.17)} \\
\implies \Pr\Big(\sum_{j\neq i} |\nu_{j,k,n}|^2 \leq (\mu + r\sigma)B\Big) \ &\geq \ \frac{r^2}{1+r^2} \qquad \text{(C.18)}
\end{aligned}$$

where $r > 0$ is a fixed number.

Now, we break the expectation in (C.16) into two parts — one with $\sum_{j\neq i}|\nu_{j,k,n}|^2 > (\mu+r\sigma)B$ and other with $\sum_{j\neq i}|\nu_{j,k,n}|^2 \leq (\mu+r\sigma)B$. We then ignore the first part to obtain another lower bound. Therefore, we now have

$$\begin{aligned}
\mathcal{C}^* &\geq \sum_{i=1}^{B}\sum_{n=1}^{N} \mathrm{E}\left\{ \frac{\max_k \log\left(1+P_{\mathsf{con}}\,\gamma_{i,k,n}\right)}{N+(\mu+r\sigma)BP_{\mathsf{con}}\beta^2 r_0^{-2\alpha}}\Big|_{\sum_{j\neq i}|\nu_{j,k,n}|^2\leq(\mu+r\sigma)B}\right\} \\
&\quad \times \Pr\Big(\sum_{j\neq i}|\nu_{j,k,n}|^2 \leq (\mu+r\sigma)B\Big) \\
&\geq \frac{\frac{r^2}{1+r^2}}{N+(\mu+r\sigma)BP_{\mathsf{con}}\beta^2 r_0^{-2\alpha}} \sum_{i=1}^{B}\sum_{n=1}^{N} \mathrm{E}\left\{ \max_k \log\left(1+P_{\mathsf{con}}\,\gamma_{i,k,n}\right)\right\} \quad \text{(C.19)} \\
&= f_{\mathsf{lo}}^{\mathsf{DN}}(r,B,N)\sum_{i,n}\mathrm{E}\left\{\log\left(1+P_{\mathsf{con}}\,\max_k \gamma_{i,k,n}\right)\right\}, \quad \text{(C.20)}
\end{aligned}$$

where (C.19) follows because $\sum_{j\neq i}|\nu_{j,k,n}|^2$ is independent of $\nu_{i,k,n}$ (and hence, independent of $\gamma_{i,k,n}$). Note that for Rayleigh fading channels, $\mu=\sigma=1$. □

Lemma 11 and earlier proved Theorem 1 show that the lower and upper bounds on $\mathcal{C}^*$ are functions of $\max_k \gamma_{i,k,n}$. To compute $\max_k \gamma_{i,k,n}$ for large K, we prove Lemma 12 and Lemma 13.

Lemma 12. *Under Rayleigh fading, i.e., $\nu_{i,k,n}\sim\mathcal{CN}(0,1)$, the CDF of $\gamma_{i,k,n}$ is given by*

$$\begin{aligned}
F_{\gamma_{i,k,n}}(\gamma) &= 1-\frac{r_0^2}{p^2}e^{-\frac{\gamma}{\beta^2 r_0^{-2\alpha}}} - \frac{1}{\alpha\beta^2p^2}\int_{\frac{\beta^2}{(p-d)^{2\alpha}}}^{\beta^2 r_0^{-2\alpha}} e^{-\frac{\gamma}{g}}\Big(\frac{g}{\beta^2}\Big)^{-1-\frac{1}{\alpha}}dg \\
&\quad + \int_{\frac{\beta^2}{(p+d)^{2\alpha}}}^{\frac{\beta^2}{(p-d)^{2\alpha}}} \exp(-\gamma/g)ds(g), \quad \text{(C.21)}
\end{aligned}$$

where $d=\sqrt{a_i^2+b_i^2}$, and

$$\begin{aligned}
s(g) &= \frac{1}{\pi p^2}\Bigg[\Big(\frac{g}{\beta^2}\Big)^{-1/\alpha}\cos^{-1}\left(\frac{d^2+\left(\frac{g}{\beta^2}\right)^{-1/\alpha}-p^2}{2d\left(\frac{g}{\beta^2}\right)^{-1/2\alpha}}\right)+p^2\cos^{-1}\left(\frac{d^2+p^2-\left(\frac{g}{\beta^2}\right)^{-1/\alpha}}{2dp}\right) \\
&\quad -\frac{1}{2}\sqrt{\Big(p+d-\Big(\frac{g}{\beta^2}\Big)^{-1/2\alpha}\Big)\Big(p+\Big(\frac{g}{\beta^2}\Big)^{-1/2\alpha}-d\Big)} \\
&\quad \times\sqrt{\Big(d+\Big(\frac{g}{\beta^2}\Big)^{-1/2\alpha}-p\Big)\Big(d+p+\Big(\frac{g}{\beta^2}\Big)^{-1/2\alpha}\Big)}\Bigg].
\end{aligned}$$

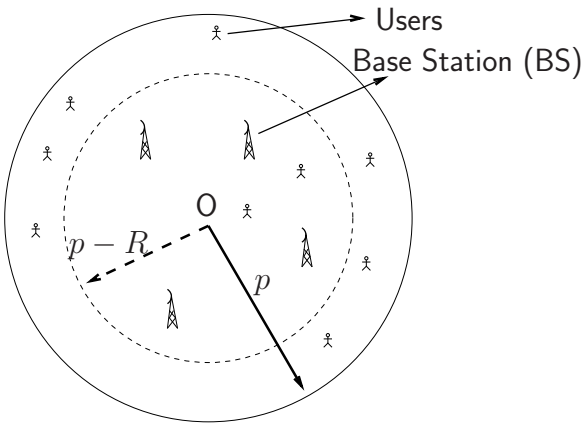

Figure C.1: OFDMA downlink system with K users and B base-stations.

Proof. We assume that the users are distributed uniformly in a circular area of radius p and there are B base-stations in that area as shown in Fig. C.1.

The probability density function of the user-coordinates (x_k, y_k) can be written as

$$f_{(x_k,y_k)}(x,y) = \begin{cases} \frac{1}{\pi p^2} & x^2 + y^2 \leq p^2 \\ 0 & \text{otherwise.} \end{cases} \tag{C.22}$$

Note that around any base-station, the users are distributed at-least within a distance R ($R > r_0$). Hence, $p - d = p - \sqrt{a_i^2 + b_i^2} \geq R > r_0$ for all i. Now,

$$\gamma_{i,k,n} = \overbrace{\Big(\underbrace{\max\Big\{r_0, \sqrt{(x_k - a_i)^2 + (y_k - b_i)^2}\Big\}}_{R_{i,k}}\Big)^{-2\alpha}}^{G_{i,k}} \beta^2 \, |\nu_{i,k,n}|^2 \tag{C.23}$$

$$= \min\Big\{r_0^{-2\alpha}, \big((x_k - a_i)^2 + (y_k - b_i)^2\big)^{-\alpha}\Big\} \beta^2 |\nu_{i,k,n}|^2. \tag{C.24}$$

We now compute the probability density function of $G_{i,k}$ ($= \beta^2 R_{i,k}^{-2\alpha}$).

$$\begin{aligned} &\Pr(G_{i,k} > g) \\ &= \Pr\left(r_0^{-2\alpha} > \frac{g}{\beta^2}\right) \times \Pr\left(\big((x_k - a_i)^2 + (y_k - b_i)^2\big)^{-\alpha} > \frac{g}{\beta^2}\right) \\ &= \Pr\left(r_0 < \left(\frac{g}{\beta^2}\right)^{-1/2\alpha}\right) \times \Pr\left(\sqrt{(x_k - a_i)^2 + (y_k - b_i)^2} < \left(\frac{g}{\beta^2}\right)^{-1/2\alpha}\right) \\ &= \begin{cases} 0 & \text{if } g \geq \beta^2 r_0^{-2\alpha} \\ \Pr\left(\sqrt{(x_k - a_i)^2 + (y_k - b_i)^2} < \left(\frac{g}{\beta^2}\right)^{-1/2\alpha}\right) & \text{otherwise.} \end{cases} \end{aligned} \tag{C.25}$$

Now, $\Pr\left(\sqrt{(x_k-a_i)^2+(y_k-b_i)^2} < \left(\frac{g}{\beta^2}\right)^{-1/2\alpha}\right)$ is basically the probability that the distance between the user k and BS i is less than $\left(\frac{g}{\beta^2}\right)^{-1/2\alpha}$. Since, the users are uniformly distributed, this probability is precisely equal to $\frac{1}{\pi p^2}$ times the intersection area of the overall area (of radius p around O) and a circle around BS i with a radius of $\left(\frac{g}{\beta^2}\right)^{-1/2\alpha}$. This is shown as the shaded region in Fig. C.2.

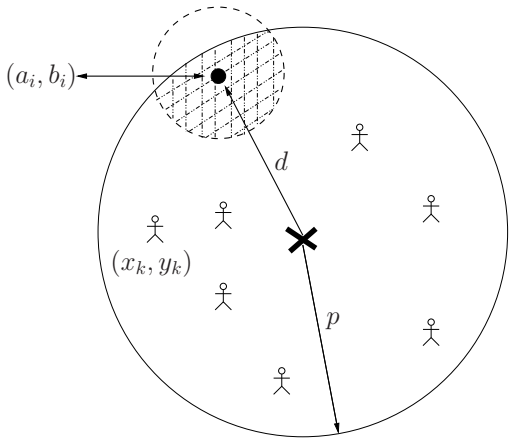

Figure C.2: System Layout. The BS i is located at a distance of d from the center with the coordinates (a_i, b_i), and the user is stationed at (x_k, y_k).

Therefore, we have:

$$\Pr(G_{i,k} > g) = \begin{cases} 1 & \text{if } \left(\frac{g}{\beta^2}\right)^{-1/2\alpha} \in (p+d, \infty) \\ s(g) & \text{if } \left(\frac{g}{\beta^2}\right)^{-1/2\alpha} \in (p-d, p+d] \\ \left(\frac{g}{\beta^2}\right)^{-1/\alpha}\frac{1}{p^2} & \text{if } \left(\frac{g}{\beta^2}\right)^{-1/2\alpha} \in (r_0, p-d] \\ 0 & \text{if } \left(\frac{g}{\beta^2}\right)^{-1/2\alpha} \in [0, r_0], \end{cases} \tag{C.26}$$

where $s(g)$ equals

$$\frac{1}{\pi p^2}\left[\left(\frac{g}{\beta^2}\right)^{-1/\alpha}\cos^{-1}\left(\frac{d^2+\left(\frac{g}{\beta^2}\right)^{-1/\alpha}-p^2}{2d\left(\frac{g}{\beta^2}\right)^{-1/2\alpha}}\right) + p^2\cos^{-1}\left(\frac{d^2+p^2-\left(\frac{g}{\beta^2}\right)^{-1/\alpha}}{2dp}\right)\right.$$
$$-\frac{1}{2}\sqrt{\left(p+d-\left(\frac{g}{\beta^2}\right)^{-1/2\alpha}\right)\left(p+\left(\frac{g}{\beta^2}\right)^{-1/2\alpha}-d\right)}$$
$$\left.\times\sqrt{\left(d+\left(\frac{g}{\beta^2}\right)^{-1/2\alpha}-p\right)\left(d+p+\left(\frac{g}{\beta^2}\right)^{-1/2\alpha}\right)}\right]. \tag{C.27}$$

The CDF of $G_{i,k}$ can now be written as

$$F_{G_{i,k}}(g) = \begin{cases} 0 & \text{if } g \in \left[0, \beta^2(p+d)^{-2\alpha}\right) \\ 1 - s(g) & \text{if } g \in \left[\beta^2(p+d)^{-2\alpha}, \beta^2(p-d)^{-2\alpha}\right) \\ 1 - \left(\frac{g}{\beta^2}\right)^{-1/\alpha} \frac{1}{p^2} & \text{if } g \in \left[\beta^2(p-d)^{-2\alpha}, \beta^2 r_0^{-2\alpha}\right) \\ 1 & \text{if } g \in \left[\beta^2 r_0^{-2\alpha}, \infty\right), \end{cases} \tag{C.28}$$

A plot of the above CDF is shown in Fig. C.3.

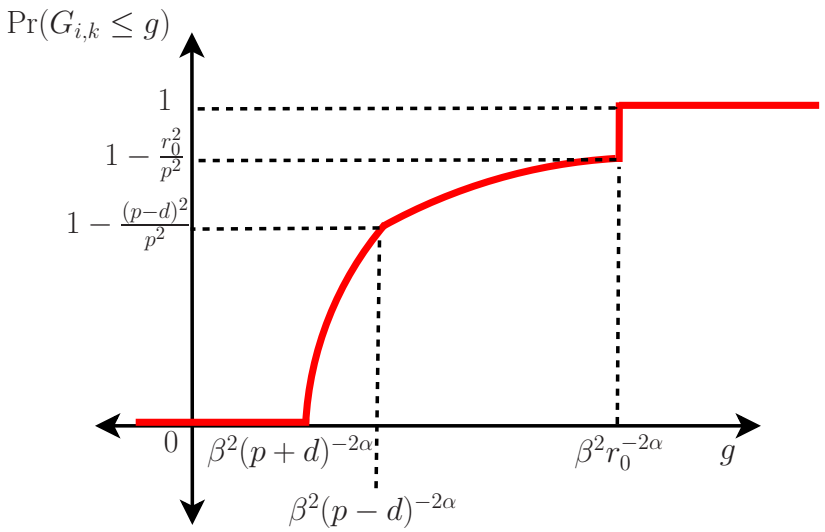

Figure C.3: Cumulative distribution function of $G_{i,k}$.

The probability density function of $G_{i,k}$ can be written as follows:

$$f_{G_{i,k}}(g) = \begin{cases} 0 & \text{if } g \in \left[0, \beta^2(p+d)^{-2\alpha}\right) \\ -\frac{ds(g)}{dg} & \text{if } g \in \left[\beta^2(p+d)^{-2\alpha}, \beta^2(p-d)^{-2\alpha}\right) \\ \frac{1}{\alpha\beta^2 p^2}\left(\frac{g}{\beta^2}\right)^{-1-1/\alpha} & \text{if } g \in \left[\beta^2(p-d)^{-2\alpha}, \beta^2 r_0^{-2\alpha}\right) \\ \frac{r_0^2}{p^2} & \text{if } g = \beta^2 r_0^{-2\alpha} \\ 0 & \text{if } g > \beta^2 r_0^{-2\alpha}, \end{cases} \tag{C.29}$$

where $\frac{ds(g)}{dg} \leq 0$. The pdf of $G_{i,k}$ has a discontinuity of the *first-kind* at $\beta^2 r_0^{-2\alpha}$ (where it takes an impulse value), and is continuous in $[\beta^2(p+d)^{-2\alpha}, \beta^2 r_0^{-2\alpha})$. At all other points, it takes the value 0.

Using (C.29), the cumulative distribution function of $\gamma_{i,k,n}$, i.e., $F_{\gamma_{i,k,n}}(\gamma)$ (when $\gamma \geq 0$) can be written as

$$\begin{aligned}
&F_{\gamma_{i,k,n}}(\gamma) \\
&= \int p\Big(|\nu_{i,k,n}|^2 \leq \frac{\gamma}{g}\Big) f_{G_{i,k}}(g)dg && \text{(C.30)} \\
&= \int \big(1 - e^{-\gamma/g}\big) f_{G_{i,k}}(g)dg && \text{(C.31)} \\
&= 1 - \int e^{-\gamma/g} f_{G_{i,k}}(g)dg && \text{(C.32)} \\
&= 1 - \frac{r_0^2}{p^2} e^{-\frac{\gamma}{\beta^2 r_0^{-2\alpha}}} - \int_{\beta^2(p-d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} e^{-\gamma/g} \frac{1}{\alpha\beta^2 p^2} \big(\frac{g}{\beta^2}\big)^{-1-1/\alpha} dg \\
&\quad + \int_{\beta^2(p+d)^{-2\alpha}}^{\beta^2(p-d)^{-2\alpha}} e^{-\gamma/g} ds(g). && \text{(C.33)}
\end{aligned}$$

□

Lemma 13. *Let $\gamma_{i,k,n}$ be a random variable with a cdf defined in Lemma 12. Then, the growth function $\frac{1-F_{\gamma_{i,k,n}}(\gamma)}{f_{\gamma_{i,k,n}}(\gamma)}$ converges to a constant $\beta^2 r_0^{-2\alpha}$ as $\gamma \to \infty$, and $\gamma_{i,k,n}$ belongs to a domain of attraction [86]. Furthermore, the cdf of $(\max_k \gamma_{i,k,n} - l_K)$ converges in distribution to a limiting random variable with a Gumbel type cdf, that is given by*

$$\exp(-e^{-x r_0^{2\alpha}/\beta^2}),\ x \in (-\infty, \infty), \tag{C.34}$$

where l_K is such that $F_{\gamma_{i,k,n}}(l_K) = 1 - 1/K$. In particular, $l_K = \beta^2 r_0^{-2\alpha} \log \frac{K r_0^2}{p^2}$.

Proof. We have from Lemma 12

$$\begin{aligned}
&F_{\gamma_{i,k,n}}(\gamma) \\
&= 1 - \frac{r_0^2}{p^2} e^{-\frac{\gamma}{\beta^2 r_0^{-2\alpha}}} - \int_{\beta^2(p-d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} e^{-\gamma/g} \frac{1}{\alpha\beta^2 p^2} \big(\frac{g}{\beta^2}\big)^{-1-1/\alpha} dg \\
&\quad + \int_{\beta^2(p+d)^{-2\alpha}}^{\beta^2(p-d)^{-2\alpha}} e^{-\gamma/g} s'(g)dg \\
&= 1 - \frac{r_0^2}{p^2} e^{-\frac{\gamma}{\beta^2 r_0^{-2\alpha}}} - \int_{\beta^2(p-d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} e^{-\gamma/g} \frac{1}{\alpha\beta^2 p^2} \big(\frac{g}{\beta^2}\big)^{-1-1/\alpha} dg \\
&\quad + s(g)e^{-\gamma/g}\Big|_{\beta^2(p+d)^{-2\alpha}}^{\beta^2(p-d)^{-2\alpha}} - \gamma \int_{\beta^2(p+d)^{-2\alpha}}^{\beta^2(p-d)^{-2\alpha}} \frac{e^{-\gamma/g} s(g)}{g^2} dg && \text{(C.35)} \\
&= 1 - \frac{r_0^2}{p^2} e^{-\frac{\gamma}{\beta^2 r_0^{-2\alpha}}} - \int_{\beta^2(p-d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} e^{-\frac{\gamma}{g}} \frac{1}{\alpha\beta^2 p^2} \big(\frac{g}{\beta^2}\big)^{-1-\frac{1}{\alpha}} dg \\
&\quad + e^{-\frac{\gamma}{\beta^2(p-d)^{-2\alpha}}} \frac{(p-d)^2}{p^2} - e^{-\frac{\gamma}{\beta^2(p+d)^{-2\alpha}}} - \gamma \int_{\beta^2(p+d)^{-2\alpha}}^{\beta^2(p-d)^{-2\alpha}} \frac{e^{-\frac{\gamma}{g}} s(g)}{g^2} dg, && \text{(C.36)}
\end{aligned}$$

where $\frac{r_0^2}{p^2} < \frac{(p-d)^2}{p^2} \leq s(g) \leq 1$ (see Fig. C.3). Now, we claim that

$$\lim_{\gamma\to\infty} \big(1 - F_{\gamma_{i,k,n}}(\gamma)\big) e^{\frac{\gamma}{\beta^2 r_0^{-2\alpha}}} = \frac{r_0^2}{p^2}. \tag{C.37}$$

It is clear that the first two terms in (C.36) contribute everything to the limit in (C.37). We will consider the rest of the terms now and show that they contribute zero towards the limit in RHS of (C.37). First, considering the 4^{th}, 5^{th}, and 6^{th} terms, we have

$$\begin{aligned}
\lim_{\gamma\to\infty} e^{\frac{\gamma}{\beta^2 r_0^{-2\alpha}}} \quad \times \quad & \left| e^{-\frac{\gamma}{\beta^2(p-d)^{-2\alpha}}} \frac{(p-d)^2}{p^2} - e^{-\frac{\gamma}{\beta^2(p+d)^{-2\alpha}}} - \gamma \int_{\beta^2(p+d)^{-2\alpha}}^{\beta^2(p-d)^{-2\alpha}} \frac{e^{-\frac{\gamma}{g}} s(g)}{g^2} dg \right| \\
\leq \quad & \lim_{\gamma\to\infty} e^{\frac{\gamma}{\beta^2 r_0^{-2\alpha}}} \left(\left| e^{-\frac{\gamma}{\beta^2(p-d)^{-2\alpha}}} \frac{(p-d)^2}{p^2} \right| + \left| e^{-\frac{\gamma}{\beta^2(p+d)^{-2\alpha}}} \right| \right. \\
& \left. + \left| \gamma \int_{\beta^2(p+d)^{-2\alpha}}^{\beta^2(p-d)^{-2\alpha}} \frac{e^{-\frac{\gamma}{g}} s(g)}{g^2} dg \right| \right) \qquad \text{(C.38)} \\
\leq \quad & \lim_{\gamma\to\infty} \frac{(p-d)^2}{p^2} e^{-\frac{\gamma}{\beta^2}((p-d)^{2\alpha} - r_0^{2\alpha})} + e^{-\frac{\gamma}{\beta^2}((p+d)^{2\alpha} - r_0^{2\alpha})} \\
& + \gamma \frac{e^{-\frac{\gamma}{\beta^2}((p-d)^{2\alpha} - r_0^{2\alpha})}}{\beta^4 (p+d)^{-4\alpha}} \qquad \text{(C.39)} \\
= \quad & 0. \qquad \text{(C.40)}
\end{aligned}$$

Now, we consider the third term in (C.36). We will show that

$$\lim_{\gamma\to\infty} \underbrace{e^{\frac{\gamma}{\beta^2 r_0^{-2\alpha}}} \times \int_{\beta^2(p-d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} e^{-\frac{\gamma}{g}} \frac{1}{\alpha\beta^2 p^2} \Big(\frac{g}{\beta^2}\Big)^{-1-\frac{1}{\alpha}} dg}_{\mathcal{T}(\gamma)} = 0. \tag{C.41}$$

Taking the first exponential term inside the integral, we have

$$\mathcal{T}(\gamma) = \int_{\beta^2(p-d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} e^{-\frac{\gamma}{g} + \gamma r_0^{2\alpha}/\beta^2} \frac{1}{\alpha\beta^2 p^2} \Big(\frac{g}{\beta^2}\Big)^{-1-\frac{1}{\alpha}} dg. \tag{C.42}$$

Substituting γ/g with x, we get

$$\mathcal{T}(\gamma) = \int_{\frac{\gamma r_0^{2\alpha}}{\beta^2}}^{\frac{\gamma(p-d)^{2\alpha}}{\beta^2}} e^{-x + \gamma r_0^{2\alpha}/\beta^2} \frac{1}{\alpha\beta^2 p^2} \Big(\frac{\gamma}{x\beta^2}\Big)^{-1-\frac{1}{\alpha}} \Big(\frac{\gamma}{x^2}\Big) dx. \tag{C.43}$$

Again substituting $x - \gamma r_0^{2\alpha}/\beta^2$ by y, we have

$$\begin{aligned}
\mathcal{T}(\gamma) &= \frac{1}{\alpha p^2}\Big(\frac{\gamma}{\beta^2}\Big)^{-\frac{1}{\alpha}} \int_0^{\gamma\frac{(p-d)^{2\alpha}-r_0^{2\alpha}}{\beta^2}} e^{-y}\Big(y+\frac{\gamma r_0^{2\alpha}}{\beta^2}\Big)^{-1+\frac{1}{\alpha}} dy && \text{(C.44)}\\
&\leq \frac{1}{\alpha p^2}\Big(\frac{\gamma}{\beta^2}\Big)^{-\frac{1}{\alpha}}\Big(\frac{\gamma r_0^{2\alpha}}{\beta^2}\Big)^{-1+\frac{1}{\alpha}} \int_0^{\frac{\left((p-d)^{2\alpha}-r_0^{2\alpha}\right)\gamma}{\beta^2}} e^{-y} dy && \text{(C.45)}\\
&= \frac{\beta^2}{\alpha\gamma p^2} r_0^{-2\alpha+2}\Big(1-e^{-\gamma\frac{(p-d)^{2\alpha}-r_0^{2\alpha}}{\beta^2}}\Big), && \text{(C.46)}
\end{aligned}$$

where, in (C.45), an upper bound is taken by putting $y = 0$ in the term $\Big(y+\frac{\gamma r_0^{2\alpha}}{\beta^2}\Big)^{-1+\frac{1}{\alpha}}$ inside the integral. Since $\mathcal{T}(\gamma)$ is positive, (C.46) shows that $\lim_{\gamma\to\infty}\mathcal{T}(\gamma) = 0$. Hence, the claim is true.

Now, after computing the derivative of $F_{\gamma_{i,k,n}}(\gamma)$ w.r.t. γ to obtain the probability density function $f_{\gamma_{i,k,n}}(\gamma)$, we have

$$\lim_{\gamma\to\infty} f_{\gamma_{i,k,n}}(\gamma) e^{\gamma r_0^{2\alpha}/\beta^2} = \frac{r_0^2}{p^2\beta^2 r_0^{-2\alpha}}. \tag{C.47}$$

We do not prove the above equation here as (C.47) is straightforward to verify (similar to the steps taken to prove (C.37)). From (C.37) and (C.47), we obtain that the growth function converges to a constant, i.e.,

$$\lim_{\gamma\to\infty} \frac{1-F_{\gamma_{i,k,n}}(\gamma)}{f_{\gamma_{i,k,n}}(\gamma)} = \beta^2 r_0^{-2\alpha}. \tag{C.48}$$

The above equation implies that $\gamma_{i,k,n}$ belongs to a *domain of maximal attraction* [86, pp. 296]. In particular, the cdf of $(\max_k \gamma_{i,k,n} - l_K)$ converges in distribution to a limiting random variable with an extreme-value cdf, that is given by [87, Definition 8.3.1]

$$\exp(-e^{-x r_0^{2\alpha}/\beta^2}),\ x \in (-\infty, \infty). \tag{C.49}$$

Here, l_K is such that $F_{\gamma_{i,\cdot,n}}(l_K) = 1 - 1/K$. Solving for l_K, we have

$$\begin{aligned}
\frac{1}{K} &= \frac{r_0^2}{p^2} e^{-\frac{l_K}{\beta^2 r_0^{-2\alpha}}} + \int_{\frac{\beta^2}{(p-d)^{2\alpha}}}^{\frac{\beta^2}{r_0^{2\alpha}}} e^{-\frac{l_K}{g}} \frac{1}{\alpha\beta^2 p^2}\Big(\frac{g}{\beta^2}\Big)^{-1-\frac{1}{\alpha}} dg\\
&\quad + \int_{\frac{\beta^2}{(p+d)^{2\alpha}}}^{\frac{\beta^2}{(p-d)^{2\alpha}}} e^{-\frac{l_K}{g}}\big(-s'(g)\big) dg
\end{aligned} \tag{C.50}$$

Substituting l_K/g by x in the first integral in RHS of (C.50) and computing an upper bound, we get

$$\begin{aligned}
\frac{1}{K} &\leq \frac{r_0^2}{p^2}e^{-\frac{l_K}{\beta^2 r_0^{-2\alpha}}} + \frac{1}{\alpha\beta^2 p^2}\int_{\frac{l_K}{\beta^2(p-d)^{-2\alpha}}}^{\frac{l_K}{\beta^2 r_0^{-2\alpha}}} e^{-x}\Big(\frac{l_K}{x\beta^2}\Big)^{-1-\frac{1}{\alpha}}\Big(\frac{-l_K}{x^2}\Big)dx \\
&\quad - e^{-\frac{l_K}{\beta^2(p-d)^{-2\alpha}}}\int_{\frac{\beta^2}{(p+d)^{2\alpha}}}^{\frac{\beta^2}{(p-d)^{2\alpha}}} s'(g)dg \qquad \text{(C.51)} \\
&= \exp\Big(-\frac{l_K}{\beta^2 r_0^{-2\alpha}}\Big)\frac{r_0^2}{p^2} + \frac{1}{\alpha p^2}\Big(\frac{l_K}{\beta^2}\Big)^{-\frac{1}{\alpha}}\int_{\frac{l_K}{\beta^2 r_0^{-2\alpha}}}^{\frac{l_K}{\beta^2(p-d)^{-2\alpha}}} e^{-x}x^{-1+\frac{1}{\alpha}}dx \\
&\quad + e^{-\frac{l_K}{\beta^2(p-d)^{-2\alpha}}}\Big(-s(g)\Big)\Big|_{g=\frac{\beta^2}{(p+d)^{2\alpha}}}^{g=\frac{\beta^2}{(p-d)^{2\alpha}}} \\
&\leq \frac{r_0^2}{p^2}e^{-\frac{l_K}{\beta^2 r_0^{-2\alpha}}} + \frac{1}{\alpha p^2}\Big(\frac{l_K}{\beta^2}\Big)^{-\frac{1}{\alpha}}\Big(\frac{l_K r_0^{2\alpha}}{\beta^2}\Big)^{-1+\frac{1}{\alpha}}\int_{\frac{l_K}{\beta^2 r_0^{-2\alpha}}}^{\frac{l_K}{\beta^2(p-d)^{-2\alpha}}} e^{-x}dx \\
&\quad + e^{-\frac{l_K}{\beta^2(p-d)^{-2\alpha}}}\Big(1-\frac{(p-d)^2}{p^2}\Big) \qquad \text{(C.52)} \\
&\leq e^{-\frac{l_K}{\beta^2 r_0^{-2\alpha}}}\frac{r_0^2}{p^2} + \frac{r_0^{2-2\alpha}}{\alpha p^2}\Big(\frac{l_K}{\beta^2}\Big)^{-1}\int_{\frac{l_K}{\beta^2 r_0^{-2\alpha}}}^{\infty} e^{-x}dx + e^{-\frac{l_K}{\beta^2(p-d)^{-2\alpha}}} \qquad \text{(C.53)} \\
&\leq e^{-\frac{l_K}{\beta^2 r_0^{-2\alpha}}}\frac{r_0^2}{p^2} + \frac{r_0^{2-2\alpha}}{\alpha p^2}\Big(\frac{l_K}{\beta^2}\Big)^{-1}e^{-\frac{l_K}{\beta^2 r_0^{-2\alpha}}} + e^{-\frac{l_K}{\beta^2(p-d)^{-2\alpha}}} \qquad \text{(C.54)} \\
&\leq e^{-\frac{l_K}{\beta^2 r_0^{-2\alpha}}}\frac{r_0^2}{p^2}\Big(1 + \frac{\beta^2 r_0^{-2\alpha}}{\alpha l_K} + \frac{p^2}{r_0^2}e^{-\frac{l_K}{\beta^2}\left((p-d)^{2\alpha}-r_0^{2\alpha}\right)}\Big) \qquad \text{(C.55)} \\
&= e^{-\frac{l_K}{\beta^2 r_0^{-2\alpha}}}\frac{r_0^2}{p^2}\Big(1+O\Big(\frac{1}{l_K}\Big)\Big). \qquad \text{(C.56)}
\end{aligned}$$

In (C.51), we substitute l_K/g by x in the first integral of (C.50), and compute an upper bound by taking the exponential term out of the second integral of (C.50). In (C.52), we note that $\frac{(p-d)^2}{p^2} \leq s(g) \leq 1$. From the above analysis, we now have

$$l_K \leq \beta^2 r_0^{-2\alpha}\log\frac{Kr_0^2}{p^2} + O\Big(\frac{1}{l_K}\Big). \qquad \text{(C.57)}$$

Now, to compute a lower bound on l_K from (C.50), we note that fact that $\frac{ds(g)}{dg} \leq 0$. Therefore,

$$\begin{aligned}
\frac{1}{K} &\geq \frac{r_0^2}{p^2}e^{-\frac{l_K}{\beta^2 r_0^{-2\alpha}}} \qquad \text{(C.58)} \\
\implies l_K &\geq \beta^2 r_0^{-2\alpha}\log\frac{Kr_0^2}{p^2}. \qquad \text{(C.59)}
\end{aligned}$$

From (C.57) and (C.59), we have $\beta^2 r_0^{-2\alpha} \log \frac{Kr_0^2}{p^2} \leq l_K \leq \beta^2 r_0^{-2\alpha} \log \frac{Kr_0^2}{p^2} + O\big(\frac{1}{\log K}\big)$. Therefore,

$$l_K \approx \beta^2 r_0^{-2\alpha} \log \frac{Kr_0^2}{p^2} \tag{C.60}$$

for large K. □

Interestingly, for a given BS i, the scaling of $\max_k \gamma_{i,k,n}$ (given by l_K in large K regime) is independent of the coordinates (a_i, b_i) and is a function of r_0, p. Now, since the growth function converges to a constant (Lemma 13), we apply [67, Theorem A.2] giving us:

$$\Pr\Big\{ l_K - \log\log K \leq \max_k \gamma_{i,k,n} \leq l_K + \log\log K \Big\} \geq 1 - O\Big(\frac{1}{\log K}\Big), \tag{C.61}$$

where $l_K = \beta^2 r_0^{-2\alpha} \log \frac{Kr_0^2}{p^2}$. Therefore,

$$\begin{aligned}
&\mathrm{E}\Big\{ \log\big(1 + P_{\mathsf{con}} \max_k \gamma_{i,k,n}\big) \Big\} \\
&\quad\leq \Pr\Big(\max_k \gamma_{i,k,n} \leq l_K + \log\log K \Big) \log(1 + P_{\mathsf{con}} l_K + P_{\mathsf{con}} \log\log K) \\
&\qquad + \Pr\Big(\max_k \gamma_{i,k,n} > l_K + \log\log K \Big) \log(1 + P_{\mathsf{con}} \beta^2 r_0^{-2\alpha} K) \qquad \text{(C.62)} \\
&\quad\leq \log(1 + P_{\mathsf{con}} l_K + P_{\mathsf{con}} \log\log K) + \log(1 + P_{\mathsf{con}} \beta^2 r_0^{-2\alpha} K) \times O\Big(\frac{1}{\log K}\Big) \\
&\quad= \log(1 + P_{\mathsf{con}} l_K) + O(1). \qquad \text{(C.63)}
\end{aligned}$$

where, in (C.62), we have used the fact that the sum-rate is bounded above by $\log(1 + P_{\mathsf{con}} \beta^2 r_0^{-2\alpha} K)$. This is because

$$\begin{aligned}
\log\big(1 + P_{\mathsf{con}} \max_k \gamma_{i,k,n}\big) \Big\} \quad &\leq \quad \log\Big(1 + P_{\mathsf{con}} \sum_k \gamma_{i,k,n}\Big) \\
&\xrightarrow{\text{w.p. } 1} \quad \log\Big(1 + P_{\mathsf{con}} K\, \mathrm{E}\{\gamma_{i,1,n}\}\Big) \qquad \text{(C.64)} \\
&\leq \quad \log\Big(1 + P_{\mathsf{con}} \beta^2 r_0^{-2\alpha} K\, \mathrm{E}\{|\nu_{i,1,n}|^2\}\Big) \\
&\leq \quad \log\Big(1 + P_{\mathsf{con}} \beta^2 r_0^{-2\alpha} K\Big). \qquad \text{(C.65)}
\end{aligned}$$

Further, from (C.61), we have

$$\begin{aligned}
&\mathrm{E}\Big\{\log\big(1+P_{\mathsf{con}}\max_k \gamma_{i,k,n}\big)\Big\}\\
&\quad\geq\ \log(1+P_{\mathsf{con}}l_K-P_{\mathsf{con}}\log\log K)\Big(1-O\Big(\frac{1}{\log K}\Big)\Big). \qquad \text{(C.66)}
\end{aligned}$$

Combining (C.63) and (C.66), we get, for large K,

$$\begin{aligned}
&BN\log(1+P_{\mathsf{con}}l_K-P_{\mathsf{con}}\log\log K)\Big(1-O\Big(\frac{1}{\log K}\Big)\Big)\\
&\leq\ \sum_{i,n}\mathrm{E}\Big\{\log\big(1+P_{\mathsf{con}}\max_k\gamma_{i,k,n}\big)\Big\} \qquad \text{(C.67)}\\
&\leq\ \big(\log(1+P_{\mathsf{con}}l_K)+O(1)\big)BN.
\end{aligned}$$

Therefore, from Lemma 11 and Theorem 1, we get

$$\begin{aligned}
&\big(\log(1+P_{\mathsf{con}}l_K)+O(1)\big)BNf_{\mathsf{lo}}^{\mathsf{DN}}(r,B,N)\\
&\leq\ \mathcal{C}^*\leq\big(\log(1+P_{\mathsf{con}}l_K)+O(1)\big)BN. \qquad \text{(C.68)}
\end{aligned}$$

This results in:

$$\begin{aligned}
\mathcal{C}^* &= O(BN\log\log K),\ \text{and}\\
\mathcal{C}^* &= \Omega(BNf_{\mathsf{lo}}^{\mathsf{DN}}(r,B,N)\log\log K). \qquad \text{(C.69)}
\end{aligned}$$

Now, to prove Corollary 1, we use the upper bound in Theorem 1 obtained via Jensen's inequality. In particular, we have

$$\begin{aligned}
\mathcal{C}^* &\leq N\sum_i\mathrm{E}\Big\{\log\Big(1+\frac{P_{\mathsf{con}}}{N}\max_{n,k}\gamma_{i,k,n}\Big)\Big\} \qquad \text{(C.70)}\\
&\leq BN\log\Big(1+\frac{P_{\mathsf{con}}}{N}l_{KN}\Big)+O(1)BN, \qquad \text{(C.71)}
\end{aligned}$$

where (C.71) follows from (C.67), and $l_{KN}=\beta^2r_0^{-2\alpha}\log\frac{KNr_0^2}{p^2}$ determines the SNR scaling of the maximum over KN i.i.d. random variables. This implies

$$\mathcal{C}^* = O\Big(BN\log\frac{\log KN}{N}\Big). \qquad \text{(C.72)}$$

Note that the above result is only true if $\frac{P_{\text{con}}}{N} l_{KN} \gg 1$ to make the approximation $\log(1+x) \approx \log x$ valid for large x.

C.3 Proof of Lemma 9 and Lemma 10

We will first find the SNR scaling laws for each of the three families of distributions — Nakagami-m, Weibull, and LogNormal. This involves deriving the domain of attraction of channel-SNR $\gamma_{i,k,n}$ for all three types of distributions. The domains of attraction are of three types - Fréchet, Weibull, and Gumbel. Let the growth function be defined as $h(\gamma) \triangleq \frac{1-F_{\gamma_{i,k,n}}(\gamma)}{f_{\gamma_{i,k,n}}(\gamma)}$. The random variable, $\gamma_{i,k,n}$, belongs to the Gumbel-type if $\lim_{\gamma\to\infty} h'(\gamma) = 0$. It turns out that all three distributions considered, i.e., Nakagami-m, Weibull, and LogNormal, belong to this category. After showing this, we find the scaling, l_K, such that $F_{\gamma_{i,k,n}}(l_K) = 1 - 1/K$. The intuition behind this choice of l_K is that the cdf of $\max_k \gamma_{i,k,n}$ is $F^K_{\gamma_{i,k,n}}(\gamma)$. For $\gamma = l_K$, we have $F^K_{\gamma_{i,k,n}}(l_K) = (1-1/K)^K \to e^{-1}$. The fact that $F^K_{\gamma_{i,k,n}}(\gamma)$ converges for a particular choice of γ gives information about the asymptotic behavior of $\max_k \gamma_{i,k,n}$.

C.3.1 Nakagami-m

In this case, $|\nu_{i,k,n}|$ is distributed according to Nakagami-(m,w) distribution. Hence, $|\nu_{i,k,n}|^2$ is distributed according to Gamma-$(m, w/m)$ distribution. The cumulative distribution function of $\gamma_{i,k,n}$, i.e., $F_{\gamma_{i,k,n}}(\gamma)$ (when $\gamma \geq 0$) is

$$
\begin{aligned}
F_{\gamma_{i,k,n}}(\gamma) &= \int p\Big(|\nu_{i,k,n}|^2 \leq \frac{\gamma}{g}\Big) f_{G_{i,k}}(g) dg && \text{(C.73)}\\
&= \int \frac{\gamma\big(m, \frac{m\gamma}{wg}\big)}{\Gamma(m)} f_{G_{i,k}}(g) dg && \text{(C.74)}\\
&= 1 - \int_{\beta^2(p+d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} \frac{\Gamma\big(m, \frac{m\gamma}{wg}\big)}{\Gamma(m)} f_{G_{i,k}}(g) dg && \text{(C.75)}
\end{aligned}
$$

where $f_{G_{i,k}}(g)$ is defined in (C.29). Now, for large γ, we can approximate (C.75) as

$$\begin{aligned} F_{\gamma_{i,k,n}}(\gamma) &\approx 1 - \frac{1}{\Gamma(m)} \int_{\beta^2(p+d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} \Big(\frac{m\gamma}{wg}\Big)^{m-1} e^{-\frac{m\gamma}{wg}} f_{G_{i,k}}(g) dg \qquad \text{(C.76)} \\ &= 1 - \frac{r_0^2}{p^2\Gamma(m)} \Big(\frac{m\gamma}{w\beta^2 r_0^{-2\alpha}}\Big)^{m-1} e^{-\frac{m\gamma}{w\beta^2 r_0^{-2\alpha}}} \\ &\quad - \frac{1}{\Gamma(m)} \int_{\beta^2(p-d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} \Big(\frac{m\gamma}{wg}\Big)^{m-1} e^{-\frac{m\gamma}{wg}} \frac{1}{\alpha\beta^2 p^2} \big(\frac{g}{\beta^2}\big)^{-1-\frac{1}{\alpha}} dg \\ &\quad + \frac{1}{\Gamma(m)} \int_{\beta^2(p+d)^{-2\alpha}}^{\beta^2(p-d)^{-2\alpha}} \Big(\frac{m\gamma}{wg}\Big)^{m-1} e^{-\frac{m\gamma}{wg}} ds(g), \qquad \text{(C.77)} \end{aligned}$$

where $f_{G_{i,k}}(g)$ is defined in (C.29). We claim that

$$\begin{aligned} &\lim_{\gamma\to\infty} \big(1 - F_{\gamma_{i,k,n}}(\gamma)\big) \gamma^{1-m} e^{\frac{m\gamma}{w\beta^2 r_0^{-2\alpha}}} \\ &= \lim_{\gamma\to\infty} \gamma^{1-m} e^{\frac{m\gamma}{w\beta^2 r_0^{-2\alpha}}} \frac{1}{\Gamma(m)} \int_{\beta^2(p+d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} \Big(\frac{m\gamma}{wg}\Big)^{m-1} e^{-\frac{m\gamma}{wg}} f_{G_{i,k}}(g) dg \qquad \text{(C.78)} \\ &= \frac{r_0^2 m^{m-1}}{p^2\Gamma(m)(w\beta^2 r_0^{-2\alpha})^{m-1}}. \qquad \text{(C.79)} \end{aligned}$$

Note that the first two terms in the RHS of (C.77) contribute everything towards the limit in (C.79). We will show that the rest of the terms contribute zero to the limit in RHS of (C.79). In particular, ignoring the constant $\Gamma(m)$, the contribution of the two integral-terms (in (C.77)) is

$$\begin{aligned} &\gamma^{1-m} e^{\frac{m\gamma}{w\beta^2 r_0^{-2\alpha}}} \Bigg(- \int_{\beta^2(p-d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} \Big(\frac{m\gamma}{wg}\Big)^{m-1} e^{-\frac{m\gamma}{wg}} \frac{1}{\alpha\beta^2 p^2} \Big(\frac{g}{\beta^2}\Big)^{-1-\frac{1}{\alpha}} dg \\ &\qquad + \int_{\beta^2(p+d)^{-2\alpha}}^{\beta^2(p-d)^{-2\alpha}} \Big(\frac{m\gamma}{wg}\Big)^{m-1} e^{-\frac{m\gamma}{wg}} ds(g) \Bigg) \\ &= \underbrace{- \int_{\beta^2(p-d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} \Big(\frac{m}{wg}\Big)^{m-1} e^{-\frac{m\gamma}{w}\left(\frac{1}{g} - \frac{1}{\beta^2 r_0^{-2\alpha}}\right)} \frac{1}{\alpha\beta^2 p^2} \Big(\frac{g}{\beta^2}\Big)^{-1-\frac{1}{\alpha}} dg}_{\mathcal{T}_1(\gamma)} \\ &\qquad + \underbrace{\int_{\beta^2(p+d)^{-2\alpha}}^{\beta^2(p-d)^{-2\alpha}} \Big(\frac{m}{wg}\Big)^{m-1} e^{-\frac{m\gamma}{w}\left(\frac{1}{g} - \frac{1}{\beta^2 r_0^{-2\alpha}}\right)} ds(g)}_{\mathcal{T}_2(\gamma)} \\ &= \mathcal{T}_1(\gamma) + \mathcal{T}_2(\gamma). \qquad \text{(C.80)} \end{aligned}$$

Now,

$$
\begin{aligned}
&|\mathcal{T}_1(\gamma)| \\
&= \left(\frac{m}{w}\right)^{m-1} \frac{\beta^{\frac{2}{\alpha}}}{\alpha p^2} \int_{\frac{\beta^2}{(p-d)^{2\alpha}}}^{\frac{\beta^2}{r_0^{2\alpha}}} g^{-m-\frac{1}{\alpha}} e^{-\frac{m\gamma}{w}\left(\frac{1}{g}-\frac{1}{\beta^2 r_0^{-2\alpha}}\right)} dg && \text{(C.81)} \\
&= \left(\frac{m}{w}\right)^{m-1} \frac{\beta^{\frac{2}{\alpha}}}{\alpha p^2} \int_{\beta^{-2} r_0^{2\alpha}}^{\beta^{-2}(p-d)^{2\alpha}} x^{m+\frac{1}{\alpha}-2} e^{-\frac{m\gamma}{w}\left(x-\beta^{-2} r_0^{2\alpha}\right)} dx && \text{(C.82)} \\
&\leq \left(\frac{m}{w}\right)^{m-1} \frac{\beta^{\frac{2}{\alpha}}}{\alpha p^2} \max\left\{\left(\frac{(p-d)^{2\alpha}}{\beta^2}\right)^{m+\frac{1}{\alpha}-2}, \left(\frac{r_0^{2\alpha}}{\beta^2}\right)^{m+\frac{1}{\alpha}-2}\right\} \\
&\quad \times \int_{\beta^{-2} r_0^{2\alpha}}^{\beta^{-2}(p-d)^{2\alpha}} e^{-\frac{m\gamma}{w}\left(x-\beta^{-2} r_0^{2\alpha}\right)} dx \\
&= \left(\frac{m}{w}\right)^{m-1} \frac{\beta^{\frac{2}{\alpha}}}{\alpha p^2} \max\left\{\left(\frac{(p-d)^{2\alpha}}{\beta^2}\right)^{m+\frac{1}{\alpha}-2}, \left(\frac{r_0^{2\alpha}}{\beta^2}\right)^{m+\frac{1}{\alpha}-2}\right\} \\
&\quad \times \frac{1 - e^{-\frac{m\gamma}{w}\left(\beta^{-2}(p-d)^{2\alpha} - \beta^{-2} r_0^{2\alpha}\right)}}{\frac{m\gamma}{w}} && \text{(C.83)} \\
&\to \ 0, \text{ as } \gamma \to \infty. && \text{(C.84)}
\end{aligned}
$$

where, in (C.82), we substituted $\frac{1}{g}$ by x. Further,

$$
\begin{aligned}
|\mathcal{T}_2(\gamma)| &= \left| \int_{\beta^2(p+d)^{-2\alpha}}^{\beta^2(p-d)^{-2\alpha}} \left(\frac{m}{wg}\right)^{m-1} e^{-\frac{m\gamma}{w}\left(\frac{1}{g}-\frac{1}{\beta^2 r_0^{-2\alpha}}\right)} ds(g) \right| && \text{(C.85)} \\
&\leq e^{-\frac{m\gamma}{w\beta^2}\left((p-d)^{2\alpha} - r_0^{2\alpha}\right)} \left| \int_{\beta^2(p+d)^{-2\alpha}}^{\beta^2(p-d)^{-2\alpha}} \left(\frac{m}{wg}\right)^{m-1} ds(g) \right| && \text{(C.86)} \\
&\to \ 0, \text{ as } \gamma \to \infty. && \text{(C.87)}
\end{aligned}
$$

Therefore, $\mathcal{T}_1(\gamma)$ and $\mathcal{T}_2(\gamma)$ have zero contribution to the RHS in (C.79), and the our claim is true. Now, from (C.77), we have

$$
\begin{aligned}
f_{\gamma_{i,k,n}}(\gamma) &= \frac{\gamma^{m-1}}{\Gamma(m)} \int_{\beta^2(p+d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} \left(\frac{m}{wg}\right)^m e^{-\frac{m\gamma}{wg}} f_{G_{i,k}}(g) dg \\
&\quad - \frac{(m-1)\gamma^{m-2}}{\Gamma(m)} \int_{\beta^2(p+d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} \left(\frac{m}{wg}\right)^{m-1} e^{-\frac{m\gamma}{wg}} f_{G_{i,k}}(g) dg
\end{aligned}
$$

Using (C.78)-(C.79), it is easy to verify that

$$
\lim_{\gamma\to\infty} f_{\gamma_{i,k,n}}(\gamma) \gamma^{1-m} e^{\frac{m\gamma}{w\beta^2 r_0^{-2\alpha}}} = \frac{r_0^2 m^m}{p^2 \Gamma(m) (w\beta^2 r_0^{-2\alpha})^m}. \qquad \text{(C.88)}
$$

From (C.79) and (C.88), we obtain that the growth function converges to a constant. In particular,

$$\lim_{\gamma\to\infty} \frac{1-F_{\gamma_{i,k,n}}(\gamma)}{f_{\gamma_{i,k,n}}(\gamma)} = \frac{w\beta^2 r_0^{-2\alpha}}{m}, \tag{C.89}$$

Hence, $\gamma_{i,k,n}$ belongs to the Gumbel-type [87, Definition 8.3.1] and $\max_k \gamma_{i,k,n} - l_K$ converges in distribution to a limiting random variable with a Gumbel-type cdf, that is given by

$$\exp(-e^{-xr_0^{2\alpha}/\beta^2}),\ x\in(-\infty,\infty), \tag{C.90}$$

where $1-F_{\gamma_{i,k,n}}(l_K) = \frac{1}{K}$. From (C.79), we have $l_K \approx \frac{w\beta^2 r_0^{-2\alpha}}{m}\log\frac{Kr_0^2 m^{m-1}}{p^2\Gamma(m)(w\beta^2 r_0^{-2\alpha})^{m-1}}$ for large K.

Now, since the growth function converges to a constant and $l_K = \Theta(\log K)$, we can use [67, Theorem 1] to obtain:

$$\Pr\Big\{l_K - \log\log K \ \leq\ \max_k \gamma_{i,k,n} \leq l_K + \log\log K\Big\} \geq 1 - O\Big(\frac{1}{\log K}\Big). \tag{C.91}$$

This is the same as (C.61). Thus, following the same analysis as in (C.62)-(C.72), we get

$$\mathcal{C}^* = O\left(BN\log\log\frac{Kr_0^2}{p^2}\right) \text{ and} \tag{C.92}$$

$$\mathcal{C}^* = BNf_{\mathsf{lo}}^{\mathsf{DN}}(r,B,N)\Omega\left(\log\log\frac{Kr_0^2}{p^2}\right). \tag{C.93}$$

Further, if $\log\frac{KN}{N} \gg 1$, then $\mathcal{C}^* = O\Big(BN\log\frac{\log\frac{KNr_0^2}{p^2}}{N}\Big)$.

C.3.2 Weibull

In this case, $|\nu_{i,k,n}|$ is distributed according to Weibull-(λ, t) distribution. Hence, $|\nu_{i,k,n}|^2$ is distributed according to Weibull-$(\lambda^2, t/2)$ distribution. We start with finding the cumulative distribution function of $\gamma_{i,k,n}$, i.e., $F_{\gamma_{i,k,n}}(\gamma)$ (when $\gamma \geq 0$) as

$$\begin{aligned}
&F_{\gamma_{i,k,n}}(\gamma) \\
&= \int p\Big(|\nu_{i,k,n}|^2 \leq \frac{\gamma}{g}\Big) f_{G_{i,k}}(g) dg && \text{(C.94)} \\
&= 1 - \int_{\beta^2(p+d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} e^{-\left(\frac{\gamma}{g\lambda^2}\right)^{t/2}} f_{G_{i,k}}(g) dg && \text{(C.95)} \\
&= 1 - \frac{r_0^2}{p^2} e^{-\left(\frac{\gamma}{\beta^2 r_0^{-2\alpha}\lambda^2}\right)^{t/2}} - \int_{\frac{\beta^2}{(p-d)^{2\alpha}}}^{\frac{\beta^2}{r_0^{2\alpha}}} \frac{e^{-\left(\frac{\gamma}{g\lambda^2}\right)^{t/2}}}{\alpha\beta^2 p^2} \Big(\frac{g}{\beta^2}\Big)^{-1-\frac{1}{\alpha}} dg \\
&\quad + \int_{\frac{\beta^2}{(p+d)^{2\alpha}}}^{\frac{\beta^2}{(p-d)^{2\alpha}}} e^{-\left(\frac{\gamma}{g\lambda^2}\right)^{t/2}} ds(g). && \text{(C.96)}
\end{aligned}$$

This case is similar to the Rayleigh distribution scenario in (C.33). Therefore, it is easy to verify that

$$\begin{aligned}
\lim_{\gamma\to\infty} \big(1 - F_{\gamma_{i,k,n}}(\gamma)\big) e^{\left(\frac{\gamma}{\beta^2 r_0^{-2\alpha}\lambda^2}\right)^{t/2}} &= \frac{r_0^2}{p^2}, \text{ and} && \text{(C.97)} \\
\lim_{\gamma\to\infty} f_{\gamma_{i,k,n}}(\gamma) \gamma^{1-t/2} e^{\left(\frac{\gamma}{\beta^2 r_0^{-2\alpha}\lambda^2}\right)^{t/2}} &= \frac{t r_0^2}{2\big(\beta^2 r_0^{-2\alpha}\lambda^2\big)^{t/2} p^2}. && \text{(C.98)}
\end{aligned}$$

Thus, the growth function $h(\gamma) = \frac{1-F_{\gamma_{i,k,n}}(\gamma)}{f_{\gamma_{i,k,n}}(\gamma)}$ can be approximated for large γ as

$$h(\gamma) \approx \frac{2\big(\beta^2 r_0^{-2\alpha}\lambda^2\big)^{t/2}}{t} \gamma^{1-t/2}. \tag{C.99}$$

Since $\lim_{\gamma\to\infty} h'(\gamma) = 0$, the limiting distribution of $\max_k \gamma_{i,k,n}$ is of Gumbel-type. Note that this is true even when $t < 1$ which refers to heavy-tail distributions. Solving for $1 - F_{\gamma_{i,k,n}}(l_K) = \frac{1}{K}$, we get

$$l_K = \beta^2 r_0^{-2\alpha} \lambda^2 \log^{\frac{2}{t}} \frac{K r_0^2}{p^2}. \tag{C.100}$$

Now, we apply the following theorem by Uzgoren.

Theorem 6 (Uzgoren)**.** *Let $x_1, \ldots, x_K$ be a sequence of i.i.d. positive random variables with continuous and strictly positive pdf $f_X(x)$ for $x > 0$ and cdf represented by $F_X(x)$. Let $h_X(x)$ be the growth function. Then, if $\lim_{x\to\infty} h'_X(x) = 0$, we have*

$$\begin{aligned}&\log\big\{ -\log F^K\big(l_K + h_X(l_K)\, u\big)\big\}\\ &= \; -u + \frac{u^2}{2!}h'_X(l_K) + \frac{u^3}{3!}\big(h_X(l_K)h''_X(l_K) - 2h'^2_X(l_K)\big) + O\bigg(\frac{e^{-u+O(u^2h'_X(l_K))}}{K}\bigg).\end{aligned}$$

Proof. See [88, Equation 19] for proof. □

The above theorem gives taylor series expansion of the limiting distribution for Gumbel-type distributions. In particular, setting $l_K = \beta^2 r_0^{-2\alpha}\lambda^2 \log^{\frac{2}{t}} \frac{Kr_0^2}{p^2}$ and $u = \log\log K$, we have $h(l_K) = O\Big(\frac{1}{\log^{-\frac{2}{t}+1} K}\Big)$, $h'(l_K) = O\Big(\frac{1}{\log K}\Big)$, $h''(l_K) = O\Big(\frac{1}{\log^{\frac{2}{t}+1} K}\Big)$, and so on. In particular, we have

$$\begin{aligned}\Pr\Big(\max_k \gamma_{i,k,n} \le l_K + h(l_K)\,\log\log K\Big) &= e^{-e^{-\log\log K + O\left(\frac{\log^2\log K}{\log K}\right)}} &\text{(C.101)}\\ &= 1 - O\Big(\frac{1}{\log K}\Big), &\text{(C.102)}\end{aligned}$$

where we have used the fact that $e^x = 1 + O(x)$ for small x. Similarly,

$$\begin{aligned}\Pr\Big(\max_k \gamma_{i,k,n} \le l_K - h(l_K)\,\log\log K\Big) &= e^{-e^{\log\log K + O\left(\frac{\log^2\log K}{\log K}\right)}} &\text{(C.103)}\\ &= e^{-\left(1+O\left(\frac{\log\log K}{\log K}\right)\right)\log K} &\text{(C.104)}\\ &= O\Big(\frac{1}{K}\Big). &\text{(C.105)}\end{aligned}$$

Subtracting (C.105) from (C.102), we get

$$\Pr\left(1 - O\left(\frac{\log\log K}{\log K}\right) < \frac{\max_k \gamma_{i,k,n}}{l_K} \le 1 + O\left(\frac{\log\log K}{\log K}\right)\right) \ge 1 - O\Big(\frac{1}{\log K}\Big) \quad \text{(C.106)}$$

Note that the above equation is the same as (C.61). Therefore, following (C.62)-(C.72), we get

$$\begin{aligned}\mathcal{C}^* &= BN\, O\Big(\log\log^{2/t}\frac{Kr_0^2}{p^2}\Big), \text{ and} &\text{(C.107)}\\ \mathcal{C}^* &= BN f_{\mathsf{lo}}^{\mathsf{DN}}(r,B,N)\,\Omega\Big(\log\log^{2/t}\frac{Kr_0^2}{p^2}\Big). &\text{(C.108)}\end{aligned}$$

Further, if $\frac{\log^{2/t} KN}{N} \gg 1$, then $\mathcal{C}^* = O\Bigg(BN \log \frac{\log^{2/t} \frac{KNr_0^2}{p^2}}{N}\Bigg)$.

C.3.3 LogNormal

In this case, $|\nu_{i,k,n}|$ is distributed according to LogNormal-(a, w) distribution. Hence, $|\nu_{i,k,n}|^2$ is distributed according to LogNormal-$(2a, 4w)$ distribution. The cumulative distribution function of $\gamma_{i,k,n}$, i.e., $F_{\gamma_{i,k,n}}(\gamma)$ (when $\gamma \geq 0$) is

$$F_{\gamma_{i,k,n}}(\gamma) = \int p\Big(|\nu_{i,k,n}|^2 \leq \frac{\gamma}{g}\Big) f_{G_{i,k}}(g) dg \tag{C.109}$$

$$= 1 - \frac{1}{2} \int_{\beta^2 (p+d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} \mathsf{erfc}\Bigg[\frac{\log \frac{\gamma}{g} - 2a}{\sqrt{8w}}\Bigg] f_{G_{i,k}}(g) dg, \tag{C.110}$$

where $\mathsf{erfc}[\cdot]$ is the complementary error function. Using the asymptotic expansion of $\mathsf{erfc}[\cdot]$, $F_{\gamma_{i,k,n}}(\gamma)$ can be approximated [89, Eq. 7.1.23] in the large γ-regime as:

$$F_{\gamma_{i,k,n}}(\gamma) \tag{C.111}$$

$$\approx 1 - \frac{1}{2} \int_{\beta^2 (p+d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} f_{G_{i,k}}(g) \frac{e^{-\left(\frac{\log \frac{\gamma}{g} - 2a}{\sqrt{8w}}\right)^2}}{\left(\frac{\log \frac{\gamma}{g} - 2a}{\sqrt{8w}}\right)\sqrt{\pi}} \sum_{m=0}^{\infty} (-1)^m \frac{(2m-1)!!}{2^m \left(\frac{\log \frac{\gamma}{g} - 2a}{\sqrt{8w}}\right)^{2m}} dg,$$

where $(2m-1)!! = 1 \times 3 \times 5 \times \ldots \times (2m-1)$. We ignore the terms $m = 1, 2, \ldots$ as the dominant term for large γ corresponds to $m = 0$. Therefore, we have

$$F_{\gamma_{i,k,n}}(\gamma)$$

$$= 1 - \sqrt{\frac{2w}{\pi}} \int_{\beta^2 (p+d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} \frac{e^{-\left(\frac{\log \frac{\gamma}{g} - 2a}{\sqrt{8w}}\right)^2}}{\log \frac{\gamma}{g} - 2a} f_{G_{i,k}}(g) dg \tag{C.112}$$

$$= 1 - \sqrt{\frac{2w}{\pi}} \frac{r_0^2}{p^2} \frac{e^{-\left(\frac{\log \frac{\gamma}{\beta^2 r_0^{-2\alpha}} - 2a}{\sqrt{8w}}\right)^2}}{\log \frac{\gamma}{\beta^2 r_0^{-2\alpha}} - 2a} - \sqrt{\frac{2w}{\pi}} \int_{\frac{\beta^2}{(p-d)^{2\alpha}}}^{\frac{\beta^2}{r_0^{2\alpha}}} \frac{1}{\alpha \beta^2 p^2} \Big(\frac{g}{\beta^2}\Big)^{-1-\frac{1}{\alpha}} \frac{e^{-\left(\frac{\log \frac{\gamma}{g} - 2a}{\sqrt{8w}}\right)^2}}{\log \frac{\gamma}{g} - 2a} dg$$

$$+ \sqrt{\frac{2w}{\pi}} \int_{\frac{\beta^2}{(p+d)^{2\alpha}}}^{\frac{\beta^2}{(p-d)^{2\alpha}}} \frac{e^{-\left(\frac{\log \frac{\gamma}{g} - 2a}{\sqrt{8w}}\right)^2}}{\log \frac{\gamma}{g} - 2a} ds(g). \tag{C.113}$$

Now, we claim that

$$\lim_{\gamma\to\infty}\big(1-F_{\gamma_{i,k,n}}(\gamma)\big)\big(\log\gamma-\log(\beta^2r_0^{-2\alpha})-2a\big)e^{\left(\frac{\log\frac{\gamma}{\beta^2r_0^{-2\alpha}}-2a}{\sqrt{8w}}\right)^2}=\frac{r_0^2}{p^2}\sqrt{\frac{2w}{\pi}}. \quad \text{(C.114)}$$

This is because the contribution of the two integrals in (C.113) towards the RHS of (C.114) is zero. The contribution of first integral, when γ is large, is

$$\begin{aligned}
&\left|\Big(\log\frac{\gamma}{\beta^2r_0^{-2\alpha}}-2a\Big)e^{\left(\frac{\log\frac{\gamma}{\beta^2r_0^{-2\alpha}}-2a}{\sqrt{8w}}\right)^2}\int_{\frac{\beta^2}{(p-d)^{2\alpha}}}^{\frac{\beta^2}{r_0^{2\alpha}}}\frac{1}{\alpha\beta^2p^2}\Big(\frac{g}{\beta^2}\Big)^{-1-\frac{1}{\alpha}}\frac{e^{-\left(\frac{\log\frac{\gamma}{g}-2a}{\sqrt{8w}}\right)^2}}{\log\frac{\gamma}{g}-2a}dg\right| \\
&\quad\leq \Big(\log\frac{\gamma}{\beta^2r_0^{-2\alpha}}-2a\Big)\frac{r_0^{-2\alpha-2}}{\alpha\beta^2p^2}\int_{\frac{\beta^2}{(p-d)^{2\alpha}}}^{\frac{\beta^2}{r_0^{2\alpha}}}\frac{e^{\left(\frac{\log\frac{\gamma}{\beta^2r_0^{-2\alpha}}-2a}{\sqrt{8w}}\right)^2-\left(\frac{\log\frac{\gamma}{g}-2a}{\sqrt{8w}}\right)^2}}{\log\frac{\gamma}{g}-2a}dg && \text{(C.115)}\\
&\quad\leq \frac{r_0^{-2\alpha-2}}{\alpha\beta^2p^2}\int_{\frac{\beta^2}{(p-d)^{2\alpha}}}^{\frac{\beta^2}{r_0^{2\alpha}}}e^{\left(\frac{\log\frac{\gamma}{\beta^2r_0^{-2\alpha}}-2a}{\sqrt{8w}}\right)^2-\left(\frac{\log\frac{\gamma}{g}-2a}{\sqrt{8w}}\right)^2}dg && \text{(C.116)}\\
&\quad\leq \frac{r_0^{-2\alpha-2}}{\alpha\beta^2p^2}\int_{\frac{\beta^2}{(p-d)^{2\alpha}}}^{\frac{\beta^2}{r_0^{2\alpha}}}e^{\frac{1}{8w}\left(\log\frac{\gamma^2}{g\beta^2r_0^{-2\alpha}}-4a\right)\log\frac{g}{\beta^2r_0^{-2\alpha}}}dg && \text{(C.117)}\\
&\quad= \frac{r_0^{-2\alpha-2}}{\alpha\beta^2p^2}\int_{\frac{\beta^2}{(p-d)^{2\alpha}}}^{\frac{\beta^2}{r_0^{2\alpha}}}\Big(\frac{g}{\beta^2r_0^{-2\alpha}}\Big)^{\frac{1}{8w}\left(\log\frac{\gamma^2}{g\beta^2r_0^{-2\alpha}}-4a\right)}dg && \text{(C.118)}\\
&\quad\leq \frac{r_0^{-2\alpha-2}}{\alpha\beta^2p^2}\int_{\frac{\beta^2}{(p-d)^{2\alpha}}}^{\frac{\beta^2}{r_0^{2\alpha}}}\Big(\frac{g}{\beta^2r_0^{-2\alpha}}\Big)^{\frac{1}{8w}\left(\log\frac{\gamma^2}{\beta^4r_0^{-4\alpha}}-4a\right)}dg && \text{(C.119)}\\
&\quad= \frac{r_0^{-2\alpha-2}}{\alpha\beta^2p^2}\frac{1}{\frac{1}{8w}\big(\log\frac{\gamma^2}{\beta^4r_0^{-4\alpha}}-4a\big)}\left(1-\Big(\frac{r_0}{p-d}\Big)^{\frac{2\alpha}{8w}\left(\log\frac{\gamma^2}{\beta^4r_0^{-4\alpha}}-4a\right)-2\alpha}\right) && \text{(C.120)}\\
&\quad\to\ 0,\ \text{as}\ \gamma\to\infty. && \text{(C.121)}
\end{aligned}$$

where in (C.115), we take an upper bound by taking the term $\left(\frac{g}{\beta^2}\right)^{-1-1/\alpha}$ out of the integral, and in (C.119), we put $g=\beta^2r_0^{-2\alpha}$ in the exponent of $\left(\frac{g}{\beta^2r_0^{-2\alpha}}\right)$ since $g\leq\beta^2r_0^{-2\alpha}$. The second integral has an exponent term that goes to zero faster than $e^{-\left(\frac{\log\frac{\gamma}{\beta^2r_0^{-2\alpha}}-2a}{\sqrt{8w}}\right)^2}\to 0$, making its contribution zero. Note that only the first two term in (C.113) contribute to the RHS in (C.114). Similar to the above analysis, it

is easy to show that

$$\lim_{\gamma\to\infty} f_{\gamma_{i,k,n}}(\gamma)\,\gamma e^{\left(\frac{\log\frac{\gamma}{\beta^2 r_0^{-2\alpha}}-2a}{\sqrt{8w}}\right)^2} = \frac{r_0^2}{p^2\sqrt{8w\pi}}. \quad \text{(C.122)}$$

Using the above equation and (C.114), we have

$$h(\gamma) = \frac{1-F_{\gamma_{i,k,n}}(\gamma)}{f_{\gamma_{i,k,n}}(\gamma)} \approx \frac{4w\gamma}{\log\gamma} \text{ for large } \gamma, \text{ and} \quad \text{(C.123)}$$

$$\lim_{\gamma\to\infty} h'(\gamma) = 0. \quad \text{(C.124)}$$

Therefore, the limiting distribution of $\max_k \gamma_{i,k,n}$ belongs to the Gumbel-type. Solving for l_K, we have

$$l_K = \beta^2 r_0^{-2\alpha} e^{\sqrt{8w\log\frac{Kr_0^2}{p^2}+\Theta(\log\log K)}}, \text{ and} \quad \text{(C.125)}$$

$h(l_K) = O\Big(\frac{l_K}{\log l_K}\Big)$, $h'(l_K) = O\Big(\frac{1}{\log l_K}\Big)$, $h''(l_K) = O\Big(\frac{1}{l_K \log l_K}\Big)$, and so on. Using Theorem 6 for $u = \log\log K$, we have

$$\Pr\Big(\max_k \gamma_{i,k,n} \le l_K + h(l_K)\,\log\log K\Big) = e^{-e^{-\log\log K + O\left(\frac{\log^2\log K}{\sqrt{\log K}}\right)}} \quad \text{(C.126)}$$

$$= 1 - O\Big(\frac{1}{\log K}\Big), \quad \text{(C.127)}$$

where we have used the fact that $e^x = 1 + O(x)$ for small x. Similarly,

$$\Pr\Big(\max_k \gamma_{i,k,n} \le l_K - h(l_K)\,\log\log K\Big) = e^{-e^{\log\log K + O\left(\frac{\log^2\log K}{\sqrt{\log K}}\right)}} \quad \text{(C.128)}$$

$$= e^{-\left(1+O\left(\frac{\log\log K}{\sqrt{\log K}}\right)\right)\log K} \quad \text{(C.129)}$$

$$= O\Big(\frac{1}{K}\Big). \quad \text{(C.130)}$$

Combining (C.127) and (C.130), we get

$$\Pr\left(l_K - c\frac{e^{\sqrt{8w\log K}}}{\log K}\log\log K < \max_k \gamma_{i,k,n} \le l_K + c\frac{e^{\sqrt{8w\log K}}}{\log K}\log\log K\right) \ge 1 - O\Big(\frac{1}{\log K}\Big), \quad \text{(C.131)}$$

where c is a constant. Now, following a similar analysis as in (C.62)-(C.72), we get

$$\max_k \gamma_{i,k,n} = \Theta\big(l_K\big) \text{ w.h.p.,} \tag{C.132}$$

$$\mathcal{C}^* = O\left(BN\sqrt{\log\frac{Kr_0^2}{p^2}}\right), \text{ and} \tag{C.133}$$

$$\mathcal{C}^* = \Omega\left(BNf_{\mathsf{lo}}^{\mathsf{DN}}(r,B,N)\sqrt{\log\frac{Kr_0^2}{p^2}}\right). \tag{C.134}$$

Further, if $\frac{e^{\sqrt{\log KN}}}{N} \gg 1$, then $\mathcal{C}^* = O\left(BN\log\frac{e^{\sqrt{\log\frac{KNr_0^2}{p^2}}}}{N}\right)$.

C.4 Proof of Theorem 4

We have $F_X(l_{T/S_1}) = 1 - \frac{S_1}{T}$, where $S_1 \in (0,T]$. Therefore, $F_{\max_t X_t}(l_{T/S_1}) = \big(1-\frac{S_1}{T}\big)^T$. This gives, for any increasing concave function $V(\cdot)$,

$$\begin{aligned}
\mathrm{E}\left\{V\big(\max_t X_t\big)\right\} &\geq \Pr\Big(\max_t X_t \geq l_{T/S_1}\Big)V\big(l_{T/S_1}\big) \\
&= \left(1-\Big(1-\frac{S_1}{T}\Big)^T\right)V\big(l_{T/S_1}\big) \qquad \text{(C.135)} \\
&\geq \big(1-e^{-S_1}\big)V\big(l_{T/S_1}\big). \qquad \text{(C.136)}
\end{aligned}$$

Setting $S_1 = \log K$ and $V(x) = \log(1+P_{\mathsf{con}}x)$, we get

$$\Big(1-\frac{1}{K}\Big)\log\big(1+P_{\mathsf{con}}l_{K/\log K}\big) \leq \mathrm{E}\Big\{\log\big(1+P_{\mathsf{con}}\max_k \gamma_{i,k,n}\big)\Big\}, \tag{C.137}$$

where

$$F_{\gamma_{i,k,n}}(l_{K/\log K}) = 1 - \frac{\log K}{K}. \tag{C.138}$$

C.5 Proof of Theorem 5

The maximum distance between a TX and user is $2p$. Therefore, we have

$$\begin{aligned}
&\mathcal{C}^*_{\mathsf{LB}} \\
&\leq \max_{\mathbf{P}\in\mathcal{P}} \sum_{i=1}^{B}\sum_{n=1}^{N} \mathrm{E}\left\{ \max_k \log\left(1+\frac{P_{i,n}\,\gamma_{i,k,n}}{1+\beta^2(2p)^{-2\alpha}\sum_{j\neq i}P_{j,n}\,|\nu_{j,k,n}|^2}\right)\right\} \\
&\leq \max_{\mathbf{P}\in\mathcal{P}} \sum_{i=1}^{B}\sum_{n=1}^{N} \mathrm{E}\left\{ \log\left(1+\max_k \frac{P_{i,n}\,\beta^2 R_{i,k}^{-2\alpha}|\nu_{i,k,n}|^2}{1+\beta^2(2p)^{-2\alpha}\sum_{j\neq i}P_{j,n}\,|\nu_{j,k,n}|^2}\right)\right\} \qquad \text{(C.139)}
\end{aligned}$$

Similarly, as a lower bound, we have (due to truncated path-loss model)

$$\begin{aligned}
&\mathcal{C}^*_{\mathsf{LB}} \\
&\geq \max_{\mathbf{P}\in\mathcal{P}} \sum_{i=1}^{B}\sum_{n=1}^{N} \mathrm{E}\left\{ \max_k \log\left(1+\frac{P_{i,n}\,\gamma_{i,k,n}}{1+\beta^2 r_0^{-2\alpha}\sum_{j\neq i}P_{j,n}\,|\nu_{j,k,n}|^2}\right)\right\} \\
&\geq \max_{\mathbf{P}\in\mathcal{P}} \sum_{i=1}^{B}\sum_{n=1}^{N} \mathrm{E}\left\{ \log\left(1+\max_k \frac{P_{i,n}\,\beta^2 R_{i,k}^{-2\alpha}|\nu_{i,k,n}|^2}{1+\beta^2 r_0^{-2\alpha}\sum_{j\neq i}P_{j,n}\,|\nu_{j,k,n}|^2}\right)\right\}. \qquad \text{(C.140)}
\end{aligned}$$

Note that the only difference in the bounds in (C.139) and (C.140) is the multiplication factor in the denominator of SINR term. In particular, the bounds can be represented as:

$$\max_{\mathbf{P}\in\mathcal{P}} \sum_{i=1}^{B}\sum_{n=1}^{N} \mathrm{E}\left\{ \log\left(1+\max_k \frac{P_{i,n}\,\beta^2 R_{i,k}^{-2\alpha}|\nu_{i,k,n}|^2}{1+\beta^2 c^{-2\alpha}\sum_{j\neq i}P_{j,n}\,|\nu_{j,k,n}|^2}\right)\right\}, \qquad \text{(C.141)}$$

where $r_0 \leq c \leq 2p$ is a constant. Defining $\mathbb{X}_{i,n}(c) \triangleq \max_k \mathbb{X}_{i,k,n}(c)$, where

$$\mathbb{X}_{i,k,n}(c) \;\triangleq\; \frac{\beta^2 R_{i,k}^{-2\alpha}|\nu_{i,k,n}|^2}{1+\beta^2 c^{-2\alpha}\sum_{j\neq i}P_{j,n}\,|\nu_{j,k,n}|^2}, \qquad \text{(C.142)}$$

the bounds can be represented as

$$\max_{\mathbf{P}\in\mathcal{P}} \sum_{i=1}^{B}\sum_{n=1}^{N} \mathrm{E}\Big\{ \log\big(1+P_{i,n}\mathbb{X}_{i,n}(c)\big)\Big\}. \qquad \text{(C.143)}$$

Let us denote $\mathbb{Y}(c) \triangleq \beta^2 c^{-2\alpha} \sum_{j \neq i} P_{j,n} |\nu_{j,k,n}|^2$. Then, we have

$$\begin{aligned} F_{\mathbb{X}_{i,k,n}(c)|R_{i,k}=r_{i,k}}(x) &= \int_{y=0}^{\infty} \Pr\left(|\nu_{i,k,n}|^2 \leq \frac{x(1+y)}{\beta^2 r_{i,k}^{-2\alpha}} \Big|_{R_{i,k}=r_{i,k}} \right) f_{\mathbb{Y}(c)}(y) dy \\ &= \int_{y=0}^{\infty} \left(1 - e^{-\frac{x(1+y)}{\beta^2 r_{i,k}^{-2\alpha}}} \right) f_{\mathbb{Y}(c)}(y) dy \\ &= 1 - \int_{y=0}^{\infty} e^{-\frac{x(1+y)}{\beta^2 r_{i,k}^{-2\alpha}}} f_{\mathbb{Y}(c)}(y) dy, \end{aligned}$$

where $F_W(x)$ denotes the value that is taken by the cdf of random variable W at x. Now, $\mathbb{Y}(c)$ has a MGF

$$M_{\mathbb{Y}(c)}(t) = \prod_{j \neq i} \frac{1}{1 - \beta^2 c^{-2\alpha} P_{j,n} t}. \tag{C.144}$$

Therefore, we have

$$\begin{aligned} F_{\mathbb{X}_{i,k,n}(c)|G_{i,k}=\beta^2 r_{i,k}^{-2\alpha}}(x) &= 1 - e^{-\frac{x}{\beta^2 r_{i,k}^{-2\alpha}}} \prod_{j \neq i} \frac{1}{1 + \beta^2 c^{-2\alpha} P_{j,n} \frac{x}{\beta^2 r_{i,k}^{-2\alpha}}} \\ &= 1 - e^{-\frac{x}{g_{i,k}}} \prod_{j \neq i} \frac{1}{1 + \frac{\beta^2 c^{-2\alpha} P_{j,n} x}{g_{i,k}}}, \end{aligned} \tag{C.145}$$

where $G_{i,k} = \beta^2 R_{i,k}^{-2\alpha}$. This gives

$$F_{\mathbb{X}_{i,k,n}}(x) = \int F_{\mathbb{X}_{i,k,n}(c)|G_{i,k}=g}(x) f_{G_{i,k}}(g) dg \tag{C.146}$$

$$\tag{C.147}$$

$$= 1 - \frac{r_0^2}{p^2} e^{-\frac{x}{g}} \prod_{j \neq i} \frac{1}{1 + \frac{\beta^2 c^{-2\alpha} P_{j,n} x}{g}} \Big|_{g=\beta^2 r_0^{-2\alpha}} \tag{C.148}$$

$$\begin{aligned} &- \int_{\beta^2 (p-d)^{-2\alpha}}^{\beta^2 r_0^{-2\alpha}} e^{-\frac{x}{g}} \left(\prod_{j \neq i} \frac{1}{1 + \frac{\beta^2 c^{-2\alpha} P_{j,n} x}{g}} \right) \frac{1}{\alpha \beta^2 p^2} \left(\frac{g}{\beta^2} \right)^{-1-1/\alpha} dg \\ &+ \int_{\beta^2 (p+d)^{-2\alpha}}^{\beta^2 (p-d)^{-2\alpha}} e^{-\frac{x}{g}} \left(\prod_{j \neq i} \frac{1}{1 + \frac{\beta^2 c^{-2\alpha} P_{j,n} x}{g}} \right) ds(g). \end{aligned} \tag{C.149}$$

At large values of x, the last two terms in the above expression are negligible compared to the second term[19]. Therefore, at large x, one can approximate

$$1 - F_{\mathbb{X}_{i,k,n}(c)}(x) \approx \frac{r_0^2}{p^2} e^{-\frac{x}{\beta^2 r_0^{-2\alpha}}} \prod_{j \neq i} \frac{1}{1 + \frac{P_{j,n} x}{c^{2\alpha} r_0^{-2\alpha}}} \text{ and} \tag{C.150}$$

$$f_{\mathbb{X}_{i,k,n}(c)}(x) \approx \frac{r_0^2}{p^2 \beta^2 r_0^{-2\alpha}} e^{-\frac{x}{\beta^2 r_0^{-2\alpha}}} \prod_{j \neq i} \frac{1}{1 + \frac{P_{j,n} x}{c^{2\alpha} r_0^{-2\alpha}}}. \tag{C.151}$$

Note that $\mathbb{X}_{i,k,n}(c)$ belongs to a domain of attraction since $\lim_{x\to\infty} \frac{1-F_{\mathbb{X}_{i,k,n}(c)}(x)}{f_{\mathbb{X}_{i,k,n}(c)}(x)} = \beta^2 r_0^{-2\alpha}$. In particular, the distribution of $\mathbb{X}_{i,n}(c) = \max_k \mathbb{X}_{i,k,n}(c)$ can be approximated by a Gumbel distribution when K is large. In particular, $\max_k \mathbb{X}_{i,k,n}(c) - l_K(c,i,n)$ converges in distribution to Gumbel-type cdf that is given by

$$\exp\{-e^{-x r_0^{2\alpha}/\beta^2}\},\ x \in (-\infty, \infty). \tag{C.152}$$

Here, $l_K(c,i,n)$ satisfies $F_{\mathbb{X}_{i,1,n}(c)}\big(l_K(c,i,n)\big) = 1 - \frac{1}{K}$.

We will now bound $\mathcal{C}^*_{\mathsf{LB}}$ via the upper and lower bounds represented by the common expression in (C.143). First, we consider the upper bound. From (C.139) and (C.143), we have

$$\mathcal{C}^*_{\mathsf{LB}} \leq \max_{\mathbf{P}\in\mathcal{P}} \sum_{i=1}^{B} \sum_{n=1}^{N} \mathrm{E}\left\{ \log\left(1 + P_{i,n} \mathbb{X}_{i,n}(2p)\right)\right\} \tag{C.153}$$

$$\leq \max_{\mathbf{P}\in\mathcal{P}} \sum_{i=1}^{B} \sum_{n=1}^{N} \log\left(1 + P_{i,n}\, \mathrm{E}\left\{\mathbb{X}_{i,n}(2p)\right\}\right), \tag{C.154}$$

where the above equation follows by Jensen's inequality. Now, we know

$$\max_k \mathbb{X}_{i,k,n}(c) - l_K(c,i,n) \xrightarrow{d} Q \tag{C.155}$$

as K tends to infinity, where $\xrightarrow{d}$ denotes convergence in distribution, and Q has a gumbel-cdf given by (C.152). Note that $\max_{k=\{1,\ldots,K\}} \mathbb{X}_{i,k,n}(c)$ is a non-decreasing

[19]Following the analysis in (C.41)-(C.46), the last but one term in (C.149) can be ignored. It is straightforward to show that the last term can be ignored at large x since the exponential term decays quickly to zero.

sequence for every realization. Hence, applying monotone convergence theorem, we have L^1 convergence. In particular,

$$\mathrm{E}\left\{ \max_k \mathbb{X}_{i,k,n}(c)\right\} \nearrow \mathrm{E}\{Q\} + l_K(c,i,n), \tag{C.156}$$

as K grows large, where $\nearrow$ denotes convergence of an increasing sequence. Therefore, $\mathrm{E}\left\{\mathbb{X}_{i,n}(2p)\right\} \leq l_K(2p,i,n) + E\{Q\}$. Noticing that $\mathrm{E}\{Q\} = \beta^2 r_0^{-2\alpha} u$, where u is the Euler-Mascheroni constant ($u \approx 0.5772$). Applying the above argument to (C.154), we have

$$\begin{aligned}
\mathcal{C}^*_{\mathsf{LB}} &\leq \max_{\mathbf{P}\in\mathcal{P}} \sum_{i=1}^{B}\sum_{n=1}^{N} \log\left(1 + P_{i,n}\left(l_K(2p,i,n) + \beta^2 r_0^{-2\alpha} u\right)\right) && \text{(C.157)}\\
&= \max_{\mathbf{P}\in\mathcal{P}} \sum_{i=1}^{B}\sum_{n=1}^{N} \log\left(1 + \left(1 + \frac{\beta^2 r_0^{-2\alpha} u}{l_K(2p,i,n)}\right) P_{i,n} l_K(2p,i,n)\right) && \text{(C.158)}\\
&\leq \max_{\mathbf{P}\in\mathcal{P}} \sum_{i=1}^{B}\sum_{n=1}^{N} \left(1 + \frac{\beta^2 r_0^{-2\alpha} u}{l_K(2p,i,n)}\right) \log\left(1 + P_{i,n} l_K(2p,i,n)\right), && \text{(C.159)}
\end{aligned}$$

where (C.159) follows from (C.158) because $\log(1+ax) \leq a\log(1+x)$ for all $x \geq 0$ and $a \geq 1$. Now, from (C.150), we know that $l = l_K(2p,i,n)$ satisfies

$$\frac{r_0^2 K}{p^2} = e^{\frac{l}{\beta^2 r_0^{-2\alpha}}} \prod_{j\neq i} \left(1 + \frac{P_{j,n} l}{(2p)^{2\alpha} r_0^{-2\alpha}}\right). \tag{C.160}$$

Note that the value of l that satisfies the above equation decreases with increase in $\{P_{j,n}$ for all $j \neq i\}$. Therefore, we can write $l_K(2p,i,n) \geq \bar{l}(2p,K)$ for all (i,n), where $l = \bar{l}(2p,K)$ is computed by solving (C.160) with $P_{j,n} = P_{\mathsf{con}}$ for all (j,n). In particular, $\bar{l}(2p,K)$ satisfies

$$\frac{r_0^2 K}{p^2} = e^{\frac{\bar{l}(2p,K)}{\beta^2 r_0^{-2\alpha}}} \left(1 + \frac{\bar{l}(2p,K) P_{\mathsf{con}}}{(2p)^{2\alpha} r_0^{-2\alpha}}\right)^{B-1}. \tag{C.161}$$

Using $l_K(2p,i,n) \geq \bar{l}(2p,K)$ in (C.159), we get

$$\mathcal{C}^*_{\mathsf{LB}} \leq \left(1 + \frac{\beta^2 r_0^{-2\alpha} u}{\bar{l}(2p,K)}\right) \max_{\mathbf{P}\in\mathcal{P}} \sum_{i=1}^{B}\sum_{n=1}^{N} \log\left(1 + P_{i,n} l_K(i,n)\right), \tag{C.162}$$

where $l = l_K(2p, i, n)$ satisfies (C.160).

We will now consider the lower bound in (C.143). The lower bound follows from Theorem 4. In particular, Using $V(\mathbb{X}_{i,n}(c)) = \log(1 + P_{i,n}\mathbb{X}_{i,n}(c))$ in Theorem 4 and taking the summation over all (i, n), the optimization problem with an objective function $\sum_{i,n} \mathrm{E}\left\{V(\mathbb{X}_{i,n})\right\}$ evaluates the lower bounds in (C.143) when $c = r_0$. Therefore, we have from Theorem 4,

$$\mathcal{C}^*_{\mathsf{LB}} \geq (1 - e^{-S_1}) \max_{\mathbf{P}\in\mathcal{P}} \sum_{i=1}^{B}\sum_{n=1}^{N} \log\big(1 + P_{i,n}\, l_{K/S_1}(r_0, i, n)\big), \tag{C.163}$$

where $S_1 \in (0, K]$ and $F_{\mathbb{X}_{i,1,n}(r_0)}\big(l_{K/S_1}(r_0, i, n)\big) = 1 - \frac{S_1}{K}$. Putting $S_1 = 1$, we have

$$\mathcal{C}^*_{\mathsf{LB}} \geq 0.63 \max_{\mathbf{P}\in\mathcal{P}} \sum_{i=1}^{B}\sum_{n=1}^{N} \log\big(1 + P_{i,n} l_K(r_0, i, n)\big), \tag{C.164}$$

Combining (C.162) and (C.164) in to one mathematical form, define the class of optimization problems as follows.

$$\begin{aligned}
&\mathsf{OP}\big(c, h(K)\big)\\
&\quad\triangleq \max_{\mathbf{P}\in\mathcal{P}} \sum_{i=1}^{B}\sum_{n=1}^{N} \log(1 + P_{i,n}x_{i,n}) && \text{(C.165)}\\
&\qquad \text{s.t. } P_{i,n} \geq 0 \;\forall\; i, n, && \text{(C.166)}\\
&\qquad\quad \sum_{n} P_{i,n} \leq P_{\mathsf{con}} \;\forall\; i, \text{ and} && \text{(C.167)}\\
&\qquad\quad \frac{r_0^2\, h(K)}{p^2} = e^{\frac{x_{i,n}}{\beta^2 r_0^{-2\alpha}}} \prod_{j\neq i}\left(1 + \frac{c^{-2\alpha}P_{j,n}x_{i,n}}{r_0^{-2\alpha}}\right) \;\forall\; i, n. && \text{(C.168)}
\end{aligned}$$

Then, we have

$$(1 - e^{-S_1})\mathsf{OP}\big(r_0, K/S_1\big) \leq \mathcal{C}^*_{\mathsf{LB}} \leq \left(1 + \frac{\beta^2 r_0^{-2\alpha} u}{\bar{l}(2p, K)}\right)\mathsf{OP}\big(2p, K\big), \tag{C.169}$$

where $S_1 \in (0, K]$, u is the Euler-Mascheroni constant and $\bar{l}(K)$ satisfies (C.161) (re-written below for brevity):

$$\frac{r_0^2 K}{p^2} = e^{\frac{\bar{l}(2p,K)}{\beta^2 r_0^{-2\alpha}}}\left(1 + \frac{\bar{l}(2p, K)P_{\mathsf{con}}}{(2p)^{2\alpha} r_0^{-2\alpha}}\right)^{B-1}. \tag{C.170}$$

C.5.1 Proof of a Property of $\mathsf{OP}\big(c, h(K)\big)$

We will now show that for positive constants c_1, c_2 $(0 < c_1 \leq c_2)$ and any increasing function $h(\cdot)$, we have

$$1 \leq \frac{\mathsf{OP}\big(c_2, h(K)\big)}{\mathsf{OP}\big(c_1, h(K)\big)} \leq \left(\frac{c_2}{c_1}\right)^{2\alpha}. \tag{C.171}$$

For any given set of powers $\{P_{i,n}$ for all $i, n\}$, let $\{x_{i,n}(c_1)$ for all $i, n\}$ be the solution to (C.168) (rewritten below for brevity) when considering the optimization problem $\mathsf{OP}\big(c_1, h(K)\big)$.

$$\frac{r_0^2\, h(K)}{p^2} = e^{\frac{x_{i,n}}{\beta^2 r_0^{-2\alpha}}} \prod_{j \neq i} \left(1 + \frac{c^{-2\alpha} P_{j,n} x_{i,n}}{r_0^{-2\alpha}}\right) \;\forall\; i, n. \tag{C.172}$$

Similarly, for the same set of powers $\{P_{i,n}$ for all $i, n\}$, let $\{x_{i,n}(c_2)$ for all $i, n\}$ be the solution to (C.172) when considering the optimization problem $\mathsf{OP}\big(c_2, h(K)\big)$. Clearly, $x_{i,n}(c_2) \geq x_{i,n}(c_1)$ since the RHS of (C.172) is a decreasing function of c. Now, we claim that

$$\left(\frac{c_2}{c_1}\right)^{2\alpha} x_{i,n}(c_1) \geq x_{i,n}(c_2) \tag{C.173}$$

for all (i, n). We know that for all (i, n)

$$\frac{r_0^2\, h(K)}{p^2} = e^{\frac{x_{i,n}(c_2)}{\beta^2 r_0^{-2\alpha}}} \prod_{j \neq i} \left(1 + \frac{c_2^{-2\alpha} P_{j,n} x_{i,n}(c_2)}{r_0^{-2\alpha}}\right) \tag{C.174}$$

Now, if we substitute $x_{i,n}(c_2)$ by any larger value, then the RHS of (C.174) will be larger than LHS of (C.174). This is because the RHS of (C.172) is an increasing function of $x_{i,n}$. Let us substitute $\left(\frac{c_2}{c_1}\right)^{2\alpha} x_{i,n}(c_1)$ instead of $x_{i,n}(c_2)$. Then, we get

$$\frac{r_0^2\, h(K)}{p^2} \geqslant e^{\frac{c_2^{2\alpha} x_{i,n}(c_1)}{c_1^{2\alpha} \beta^2 r_0^{-2\alpha}}} \prod_{j \neq i} \left(1 + \frac{c_1^{-2\alpha} P_{j,n} x_{i,n}(c_1)}{r_0^{-2\alpha}}\right), \tag{C.175}$$

where the actual inequality will be determined later. Since $\{x_{i,n}(c_1) \text{ for all } i,n\}$ is the solution to (C.172) when considering the optimization problem $\mathsf{OP}\big(c_1, h(K)\big)$, we also have

$$\frac{r_0^2\, h(K)}{p^2} = e^{\frac{x_{i,n}(c_1)}{\beta^2 r_0^{-2\alpha}}} \prod_{j\neq i} \left(1 + \frac{c_1^{-2\alpha} P_{j,n} x_{i,n}(c_1)}{r_0^{-2\alpha}}\right) \tag{C.176}$$

Dividing (C.175) by (C.176) and taking logarithm of both sides, we get

$$0 \geqslant \left(\frac{c_2^{2\alpha}}{c_1^{2\alpha}} - 1\right) \frac{x_{i,n}(c_1)}{\beta^2 r_0^{-2\alpha}}, \tag{C.177}$$

Since $c_2 \geq c_1$, we have in (C.175)

$$\frac{r_0^2\, h(K)}{p^2} \leq e^{\frac{c_2^{2\alpha} x_{i,n}(c_1)}{c_1^{2\alpha}\beta^2 r_0^{-2\alpha}}} \prod_{j\neq i} \left(1 + \frac{c_1^{-2\alpha} P_{j,n} x_{i,n}(c_1)}{r_0^{-2\alpha}}\right), \tag{C.178}$$

Therefore, $\left(\frac{c_2}{c_1}\right)^{2\alpha} x_{i,n}(c_1) \geq x_{i,n}(c_2)$ for all (i,n). Using this relation and the fact that $x_{i,n}(c_2) \geq x_{i,n}(c_1)$, we have

$$\log(1 + P_{i,n} x_{i,n}(c_1)) \leq \log(1 + P_{i,n} x_{i,n}(c_2)) \leq \log\left(1 + \left(\frac{c_2}{c_1}\right)^{2\alpha} P_{i,n} x_{i,n}(c_1)\right). \tag{C.179}$$

Also note that $\log(1 + ax) \leq a \log(1 + x)$ for all $x \geq 0$ and $a \geq 1$. Therefore, we have

$$\log\left(1 + \left(\frac{c_2}{c_1}\right)^{2\alpha} P_{i,n} x_{i,n}(c_1)\right) \leq \left(\frac{c_2}{c_1}\right)^{2\alpha} \log\big(1 + P_{i,n} x_{i,n}(c_1)\big). \tag{C.180}$$

Combining (C.179) and (C.180), we have for every (i,n)

$$1 \leq \frac{\log(1 + P_{i,n} x_{i,n}(c_2))}{\log(1 + P_{i,n} x_{i,n}(c_1))} \leq \left(\frac{c_2}{c_1}\right)^{2\alpha}. \tag{C.181}$$

Putting the above equation in $\mathsf{OP}\big(c, h(K)\big)$ (see (C.165)-(C.168)), we have

$$1 \leq \frac{\mathsf{OP}\big(c_2, h(K)\big)}{\mathsf{OP}\big(c_1, h(K)\big)} \leq \left(\frac{c_2}{c_1}\right)^{2\alpha}. \tag{C.182}$$

Bibliography

[1] A. Goldsmith, *Wirless Communications*. New York: Cambridge University Press, 2005.

[2] A. Goldsmith and S. Chua, "Variable rate variable power M-QAM for fading channels," *IEEE Trans. Commun.*, vol. 45, pp. 1218–1230, Oct. 1997.

[3] D. L. Goeckel, "Adaptive coding for time-varying channels using outdated fading estimates," *IEEE Trans. Commun.*, vol. 47, pp. 844–855, June 1999.

[4] K. Balachandran, S. R. Kadaba, and S. Nanda, "Channel quality estimation and rate adaptation for cellular mobile radio," *IEEE J. Select. Areas In Commun.*, vol. 17, pp. 1244–1256, July 1999.

[5] G. Holland, N. Vaidya, and P. Bahl, "A rate adaptive MAC protocol for multi-hop wireless networks," in *Proc. ACM Internat. Conf. on Mobile Computing and Networking*, vol. 5, pp. 3246–3250, 2001.

[6] B. Sadegi, V. Kanodia, A. Sabharwal, and E. Knightly, "Opportunistic media access for multirate ad hoc networks," in *Proc. ACM Internat. Conf. on Mobile Computing and Networking*, vol. 5, (Atlanta, GA), pp. 3246–3250, 2001.

[7] J. C. Bicket, "Bit-rate selection in wireless networks," Master's thesis, Massachusetts Institute of Technology, Feb 2005.

[8] S. Wong, H. Yang, S. Lu, and V. Bharghavan, "Robust rate adaptation for 802.11 wireless networks," in *Proc. ACM Internat. Conf. on Mobile Computing and Networking*, 2006.

[9] M. Rice and S. B. Wicker, "Adaptive error control for slowly varying channels," *IEEE Trans. Commun.*, vol. 42, pp. 917–925, Feb./Mar./Apr. 1994.

[10] Y.-D. Yao, "An effective go-back-N ARQ scheme for variable-error-rate channels," *IEEE Trans. Commun.*, vol. 43, pp. 20–23, Jan. 1995.

[11] S. S. Chakraborty, M. Liinabarja, and E. Yli-Juuti, "An adaptive ARQ scheme with packet combining for time varying channels," *IEEE Commun. Letters*, vol. 3, pp. 52–54, Feb. 1999.

[12] S. Choi and K. G. Shin, "A class of hybrid ARQ schemes for wireless links," *IEEE Trans. Veh. Tech.*, vol. 50, pp. 777–790, May 2001.

[13] H. Minn, M. Zeng, and V. K. Bhargava, "On ARQ scheme with adaptive error control," *IEEE Trans. Veh. Tech.*, vol. 50, pp. 1426–1436, Nov. 2001.

[14] A. K. Karmokar, D. V. Djonin, and V. K. Bhargava, "POMDP-based coding rate adaptation for Type-I hybrid ARQ systems over fading channels with memory," *IEEE Trans. Wireless Commun.*, vol. 5, pp. 3512–3523, Dec. 2006.

[15] D. V. Djonin, A. K. Karmokar, and V. K. Bhargava, "Joint rate and power adaptation for type-I hybrid ARQ systems over correlated fading channels under different buffer-cost constraints," *IEEE Trans. Veh. Tech.*, vol. 57, pp. 421–435, Jan. 2008.

[16] G. E. Monahan, "A survey of partially observable Markov decision processes: Theory, models, and algorithms," *Management Science*, vol. 28, pp. 1–16, Jan. 1982.

[17] C. C. Tan and N. C. Beaulieu, "On first-order Markov modeling for the Rayleigh fading channel," *IEEE Trans. Commun.*, vol. 48, pp. 2032–2040, Dec. 2000.

[18] J. G. Proakis, *Digital Communications*. New York: McGraw-Hill, 3rd ed., 1995.

[19] D. Bertsekas, *Dynamic Programming and Optimal Control*. Athena Scientific, 2nd ed., 2000.

[20] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of Markov decision processes," *Mathematics of Operations Research*, vol. 12, no. 3, pp. 441–450, 1987.

[21] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer, 2nd ed., 1994.

[22] I. Abou-Fayal, M. Medard, and U. Madhow, "Binary adaptive coded pilot symbol assisted modulation over rayleigh fading channels without feedback," *IEEE Trans. Commun.*, vol. 53, pp. 1036–1046, June 2008.

[23] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston: Kluwer Academic Publishers, 1992.

[24] P. Sadeghi, R. Kennedy, P. Rapajic, and R. Shams, "Finite-state markov modeling of fading channels - a survey of principles and applications," *IEEE Signal Processing Mag.*, vol. 25, pp. 57–80, Sept. 2008.

[25] G. Song and Y. Li, "Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks," *IEEE Commun. Mag.*, vol. 43, pp. 127 – 134, Dec. 2005.

[26] M. Falkner, M. Devetsikiotis, and I. Lambadaris, "An overview of pricing concepts for broadband IP networks," *IEEE Commun. Surveys Tutorials*, vol. 3, pp. 2–13, second quarter 2000.

[27] L. A. DaSilva, "Pricing for QoS-enabled networks: A survey," *IEEE Commun. Surveys Tutorials*, vol. 3, pp. 2–8, second quarter 2000.

[28] S. Shenker, "Fundamental design issues for the future internet," *IEEE J. Select. Areas In Commun.*, vol. 13, pp. 1176–1188, Sept. 1995.

[29] "IEEE 802.16-2005 and IEEE Std 802.16-2004/Corl-2005," http://www.ieee802.org/16.

[30] "Overview of 3GPP Release 8 V0.2.2 (2011-01)," http://www.3gpp.org/ftp/Information/WORK_PLAN/Description_Releases/.

[31] J. Huang, V. Subramanian, R. Agrawal, and R. Berry, "Downlink scheduling and resource allocation for OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 8, pp. 288–296, Jan. 2009.

[32] "3GPP Technical Specifications (Release 10) 3GPP TS 25.101 V10.1.0," http://www.3gpp.org/article/umts.

[33] I. C. Wong and B. L. Evans, "Optimal resource allocation in the OFDMA downlink with imperfect channel knowledge," *IEEE Trans. Commun.*, vol. 57, pp. 232–241, Jan. 2009.

[34] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks—Parts I and II," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 614–634, Mar. 2005.

[35] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit and power allocation," *IEEE J. Select. Areas In Commun.*, vol. 17, pp. 1747–1758, Oct. 1999.

[36] T. J. Willink and P. H. Wittke, "Optimization and performance evaluation of multicarrier transmission," *IEEE Trans. Inform. Theory*, vol. 43, pp. 426–440, Mar. 1997.

[37] L. M. C. Hoo, B. Halder, J. Tellado, and J. M. Cioffi, "Multiuser transmit optimization for multicarrier broadcast channels: Asymptotic FDMA capacity region and algorithms," *IEEE Trans. Commun.*, vol. 52, pp. 922–930, Jun. 2004.

[38] I. Wong and B. Evans, "Optimal downlink OFDMA resource allocation with linear complexity to maximize ergodic rates," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 962 –971, Mar. 2008.

[39] K. Seong, M. Mohseni, and J. M. Cioffi, "Optimal resource allocation for OFDMA downlink systems," *Proc. IEEE Int. Symposium Inform. Theory*, pp. 1394–1398, Jul. 2006.

[40] A. Ahmad and M. Assaad, "Margin adaptive resource allocation in downlink OFDMA system with outdated channel state information," *Proc. IEEE Int. Symposium Personal Indoor Mobile Radio Commun.*, pp. 1868–1872, 2009.

[41] D. Hui and V. Lau, "Design and analysis of delay-sensitive cross-layer OFDMA systems with outdated CSIT," *IEEE Trans. Wireless Commun.*, vol. 8, pp. 3484 –3491, July 2009.

[42] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[43] J. Mitchell and E. K. Lee, "Branch-and-bound methods for integer programming," in *Encyclopedia of Optimization* (A. Floudas and P. M. Pardalos, eds.), vol. 2, ch. 4, pp. 509–519, The Netherlands: Kluwer Academic Publishers, 2001.

[44] J. Jang and K. B. Lee, "Transmit power adaptation for multiuser OFDM systems," *IEEE J. Select. Areas In Commun.*, vol. 21, pp. 171–178, Feb. 2003.

[45] M. Bénichou, J. M. Gauthier, P. Girodet, G. Hentges, G. Ribière, and O. Vincent, "Experiments in mixed-integer linear programming," *Mathematical Programming*, vol. 1, pp. 76–94, 1971.

[46] A. Lodi, "Mixed integer programming computation," in *50 Years of Integer Programming: 1958-2008* (M. Jünger et al., ed.), pp. 619–645, Berlin-Heidelberg: Springer-Verlag, 2010.

[47] G. Cornuejols, M. Conforti, and G. Zambelli, "Polyhedral approaches to mixed integer programming," in *50 Years of Integer Programming: 1958-2008* (M. Jünger et al., ed.), pp. 343–385, Berlin-Heidelberg: Springer-Verlag, 2010.

[48] R. C. Jeroslow, "There cannot be any algorithm for integer programming with quadratic constraints," *Operations Research*, vol. 21, no. 1, pp. 221–224, 1973.

[49] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 1982.

[50] "UTRA-UTRAN Long Term Evolution (LTE) and 3GPP System Architecture Evolution (SAE) 3GPP LTA Paper," *3GPP*. ftp://ftp.3gpp.org/Inbox/2008_web_files/LTA_Paper.pdf.

[51] J. G. Proakis, *Digital Communications*. New York: McGraw-Hill, 5th ed., 2008.

[52] J.-J. van de Beek, O. Edfors, M. Sandell, S. Wilson, and P. Borjesson, "On channel estimation in OFDM systems," in *Proc. IEEE Veh. Tech. Conf.*, vol. 2, pp. 815–819, 1995.

[53] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 2nd ed., 1994.

[54] I. C. Wong and B. L. Evans, "Optimal OFDMA resource allocation with linear complexity to maximize ergodic weighted sum capacity," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 3, pp. III–601 –III–604, 15-20 Apr. 2007.

[55] D. J. Love, R. W. Heath, V. K. N. Lau, D. Gesbert, B. D. Rao, and M. Andrews, "An overview of limited feedback in wireless communications systems," *IEEE J. Select. Areas In Commun.*, vol. 26, pp. 1341–1365, Oct. 2008.

[56] D. P. Bertsekas and R. G. Gallager, *Data Networks*. Prentice Hall, 2nd ed., 1992.

[57] A. K. Karmokar, D. V. Djonin, and V. K. Bhargava, "POMDP-based coding rate adaptation for type-I hybrid ARQ systems over fading channels with memory," *IEEE Trans. Wireless Commun.*, vol. 5, pp. 3512–3523, Dec. 2006.

[58] R. Aggarwal, P. Schniter, and C. E. Koksal, "Rate adaptation via link-layer feedback for goodput maximization over a time-varying channel," *IEEE Trans. Wireless Commun.*, vol. 8, pp. 4276–4285, Aug. 2009.

[59] G. E. Monahan, "A survey of partially observable Markov decision processes: Theory, models, and algorithms," *Management Science*, vol. 28, pp. 1–16, Jan. 1982.

[60] M. A. Haleem and R. Chandramouli, "Adaptive downlink scheduling and rate selection: a cross-layer design," *IEEE J. Select. Areas In Commun.*, vol. 23, pp. 1287–1297, Jun. 2005.

[61] R. Wang and V. K. N. Lau, "Robust optimal cross-layer designs for TDD-OFDMA systems with imperfect CSIT and unknown interference – state-space approach based on 1-bit ACKNAK feedbacks," *IEEE Trans. Commun.*, vol. 56, pp. 754–761, May 2008.

[62] Z. K. M. Ho, V. K. N. Lau, and R. S.-K. Cheng, "Cross-layer design of FDD-OFDM systems based on ACK/NAK feedbacks," *IEEE Trans. Inform. Theory*, vol. 55, pp. 4568–4584, Oct. 2009.

[63] Y. J. Zhang and S. C. Liew, "Proportional fairness in multi-channel multi-rate wireless networks–Part II: The case of time-varying channels with application to OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 3457–3467, Sept. 2008.

[64] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of Markov decision processes," *Math. Operations Res.*, vol. 12, pp. 441–450, Aug. 1987.

[65] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Processing*, vol. 50, pp. 174–188, Feb. 2002.

[66] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: a survey," *IEEE Commun. Mag.*, vol. 46, pp. 59–67, Sept. 2008.

[67] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Trans. Inform. Theory*, vol. 51, pp. 506–522, Feb. 2005.

[68] P. Gupta and P. Kumar, "The capacity of wireless networks," *IEEE Trans. Inform. Theory*, vol. 46, pp. 388–404, Mar. 2000.

[69] L.-L. Xie and P. Kumar, "A network information theory for wireless communication: scaling laws and optimal operation," *IEEE Trans. Inform. Theory*, vol. 50, pp. 748–767, May 2004.

[70] O. Leveque and I. Telatar, "Information-theoretic upper bounds on the capacity of large extended ad hoc wireless networks," *IEEE Trans. Inform. Theory*, vol. 51, pp. 858–865, Mar. 2005.

[71] M. Franceschetti, O. Dousse, D. N. C. Tse, and P. Thiran, "Closing the gap in the capacity of wireless networks via percolation theory," *IEEE Trans. Inform. Theory*, vol. 53, pp. 1009–1018, Mar. 2007.

[72] M. Grossglauser and D. Tse, "Mobility increases the capacity of ad-hoc wireless networks," in *INFOCOM 2001*, vol. 3, pp. 1360–1369, 2001.

[73] S. Kulkarni and P. Viswanath, "A deterministic approach to throughput scaling in wireless networks," *IEEE Trans. Inform. Theory*, vol. 50, pp. 1041–1049, Jun. 2004.

[74] M. Ebrahimi, M. Maddah-Ali, and A. Khandani, "Throughput scaling laws for wireless networks with fading channels," *IEEE Trans. Inform. Theory*, vol. 53, pp. 4250–4254, Nov. 2007.

[75] P. Sripathi and J. Lehnert, "A throughput scaling law for a class of wireless relay networks," in *Proc. Asilomar Conf. Signals, Systems and Computers*, vol. 2, pp. 1333–1337, Nov. 2004.

[76] W. Choi and J. Andrews, "The capacity gain from intercell scheduling in multi-antenna systems," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 714–725, Feb. 2008.

[77] M. Kountouris and J. Andrews, "Throughput scaling laws for wireless ad hoc networks with relay selection," in *Proc. IEEE Veh. Tech. Conf.*, pp. 1–5, Apr. 2009.

[78] D. Gesbert and M. Kountouris, "Rate scaling laws in multicell networks under distributed power control and user scheduling," *IEEE Trans. Inform. Theory*, vol. 57, pp. 234–244, Jan. 2011.

[79] I. Sohn, J. Andrews, and K. B. Lee, "Capacity scaling of MIMO broadcast channels with random user distribution," in *Proc. IEEE Int. Symposium Inform. Theory*, pp. 2133–2137, Jun. 2010.

[80] J. Leinonen, J. Hamalainen, and M. Juntti, "Performance analysis of downlink ofdma resource allocation with limited feedback," *Wireless Communications, IEEE Transactions on*, vol. 8, pp. 2927 –2937, june 2009.

[81] M. Vu, "Miso capacity with per-antenna power constraint," *IEEE Trans. Commun.*, vol. 59, pp. 1268–1274, May 2011.

[82] M. Andrews, V. Capdevielle, A. Feki, and P. Gupta, "Autonomous spectrum sharing for mixed lte femto and macro cells deployments," in *INFOCOM*, pp. 1–5, Mar. 2010.

[83] B. Hassibi and T. L. Marzetta, "Multiple-antennas and isotropically random unitary inputs: The received signal density in closed form," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1473–1484, Jun. 2002.

[84] J. A. Fridy, *Introductory analysis: The theory of calculus*. Gulf Professional Publishing, 2000.

[85] H. Everett, "Generalized lagrange multiplier method for solving problems of optimum allocation of resources," *Operations Research*, vol. 11, pp. 399–417, May-June 1963.

[86] H. A. David and H. N. Nagaraja, *Order Statistics*. New York: John Wiley and Sons, 3rd ed., 2003.

[87] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *A First Course in Order Statistics*. SIAM-Society for Industrial and Applied Mathematics, 2008.

[88] N. T. Uzgoren, *"The asymptotic developement of the distribution of the extreme values of a sample" in Studies in Mathematics and Mechanics Presented to Richard von Mises*. New York: Academic, 1954.

[89] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*. 9th printing, Dover Publications, 1970.