Opportunistic Beamforming in Wireless Networks:

Optimal Selective Feedback Policies and the Feedback-Capacity Tradeoff

Tharaka Samarasinghe

Submitted in total fulfilment of the requirements of the degree of

Doctor of Philosophy

Department of Electrical and Electronic Engineering
THE UNIVERSITY OF MELBOURNE

December 2012

Abstract

OPPORTUNISTIC beamforming is a reduced feedback communication strategy for vector broadcast channels which only requires partial channel state information (CSI) at the base station for its operation. Although reducing feedback, this strategy, in its plain implementations, displays a linear growth in the feedback load with the total number of users in the system, which can be an onerous requirement for large systems. This impracticality motivates the use of selective feedback techniques in which only the users with good channels are allowed to feed back. In this thesis, we focus on a more stringent but practical finite limit on the feedback load, and we study the structure of the sum-rate maximizing decentralized selective feedback policies, and how the resulting sum-rate compare to the sum-rate without any user selection.

Firstly, we set up the problem of finding the structure of downlink sum-rate maximizing selective decentralized feedback policies for opportunistic beamforming under finite feedback constraints on the average number of mobile users feeding back. We show that any sum-rate maximizing selective decentralized feedback policy must be a threshold feedback policy. This result holds for most practical fading channel models.

Then, the resulting optimum threshold selection problem is analyzed in detail. This is a non-convex optimization problem over finite dimensional Euclidean spaces. By utilizing the theory of majorization, an underlying Schur-concave structure in the sum-rate function is identified, and the sufficient conditions for the optimality of homogenous threshold feedback policies are obtained. Applications of these results are illustrated for well known fading channel models such as Rayleigh, Nakagami and Rician fading channels, along with various engineering and design insights. Rather surprisingly, it is shown that using the same threshold value at all mobile users is not always a rate-wise

optimal feedback strategy, even for a network with statistically identical mobile users. For the Rayleigh fading channel model, on the other hand, homogenous threshold feedback policies are proven to be rate-wise optimal if multiple orthonormal data carrying beams are used to communicate with multiple mobile users simultaneously.

Having established the optimality of a homogenous threshold for a multi-beam Rayleigh fading environment, we then analyze the derivation of these optimum policies which ensure a $O(1)$ feedback constraint as the number of users grows large. Starting with a set of statistically identical users, we obtain the tradeoff curve tracing the Pareto optimal boundary between feasible and infeasible feedback-capacity pairs for opportunistic beamforming. Any point on this tradeoff curve can be obtained by means of the derived threshold feedback policies, which are rate-wise optimal. We further show to what extent the $O(1)$ feedback constraint must be relaxed to achieve the same sum-rate scaling as with full CSI. Extensions of these results to heterogeneous communication environments in which different users experience non-identical path-loss gains are also provided. We also show how threshold feedback policies can be used to provide fairness in a heterogeneous system, while simultaneously achieving optimal capacity scaling. Although the results in this analysis are asymptotic in the sense that they are derived by letting the number of users grow large, they provide promising performance results in finite size systems.

Finally, we focus on the structure of the optimal homogenous threshold feedback policy that maximizes the ergodic downlink sum-rate under a peak feedback load constraint, which we model by using a multi-packet reception model for the uplink. We solve the resulting quasi-convex optimization problem by obtaining a formula for the sum-rate maximizing threshold level. While providing insights on the implications of our results in practical systems, we also obtain the Pareto optimal boundary between feasible and infeasible feedback-capacity pairs under peak feedback load constraints.

Declaration

This is to certify that

1. the thesis comprises only my original work towards the PhD,
2. due acknowledgement has been made in the text to all other material used,
3. the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Tharaka Samarasinghe, December 2012

Acknowledgements

First and foremost, I would like to thank my supervisors Prof. Jamie Evans, Dr. Hazer Inaltekin, and Dr. Tansu Alpcan, for their guidance and support. I thank them for sharing their expertise, and specially for making time in their busy schedules to help me with my research. I would like to specially thank Prof. Evans and Dr. Inaltekin for continuously providing me support even after they left The University of Melbourne. This thesis would not have been possible without your support, and working under your supervision was a privilege.

I would like to thank my university, The University of Melbourne, for providing me financial support, great facilities, and a good work environment to conduct my research. My gratitude also goes to the Australian Government for their contribution to my tuition fees. I would like to also thank The University of Moratuwa, Sri Lanka, the institution at which I did my undergraduate degree, for providing me a sound technical foundation to carry out post-graduate studies and research.

In addition to my supervisors, it is a pleasure to also thank the following staff of the wireless communications research group for their advice and support - Dr. Brian Krongold, Prof. Subhrakanti Dey, Prof. Stephen Hanly, Dr. Sibi Raj Bhaskaran, Dr. Randa Zakhour, and Dr. Vasanthan Raghavan. Special thanks goes to Dr. Krongold who was one of my PhD advisory panel members. I would also like to thank Jackie Brisonette, Nasrin Hashemi, Angie Laoutzmis, and Anne Lopes who have helped me in various administrative activities related to my PhD.

I would also like to thank my fellow PhD students who made my time at the university very enjoyable. The following friends, who are all University of Melbourne students, have helped me one way or the other, and for that they deserve my gratitude - Athipat

Limmanee, Ehsan Nekouei, Rusdha Muharar, Chih-Hong Wang, Jack Hsu, Meng Wang, Yuan Yuan He, Rajitha Senanayake, Dinuka Kudavithana, Xiaoxi Guo, Liang Chen, Yizhuo Yang, Nader Ghasemi, Senaka Samarasekera, Duleepa Jayasundara, Sachintha Karunaratne, Chamil Jayasundara, Manoj Liyanage, Chathurika Abeyrathne, Imali Dias, and Banie Abeywickrama. I must also thank Gamini Ranasinghe, Kasun Wedage, Asanka De Silva, Chamila Kaduruwewa, Yashika Wijayawickrama, Chinthana Gunawickrama and Danisha Goonasekara for their company, for helping me out when I first moved in to Melbourne from Sri Lanka, and for also making my weekends and holidays very enjoyable.

Last but not least, I would like to express my deepest gratitude to my loving parents Chandana and Yamuna Samarasinghe, my sister Savini, and my wife Rasara, for their support, love and care. This thesis is dedicated to you.

To my loving parents and wife, for everything.

Contents

List of Figures

Chapter 1
Introduction

1.1 Multiple-input Multiple-output

IMPROVING rates of data communication over wireless links to meet the insatiable demand for multimedia traffic has become one of the pressing concerns for wireless telecommunications technology [7]. The upcoming fourth generation (4G) technology, which is also known as the *Long Term Evolution* (LTE), is envisioned to achieve data rates as high as 1 Gbit/s for users with low mobility, and 100 Mbit/s for users with high mobility [1, 12, 33]. Multiple-input multiple-output (MIMO) systems featuring multiple transmit and receive antennas at the ends of a wireless link [8, 91] marked a paradigm change from scalar communication to the vector one, and sparked great research interest in the past decade as a methodology which can facilitate such high data rates. The progress in research urged many commercial entities to develop multiple-antenna technologies for wireless networks, and today, these technologies have emerged as strong candidates in fulfilling the data rate requirements set for the next generation networks.

One of the main benefits of MIMO is the *power gain* achieved by using multiple transmit antennas. The multiple antennas also create multiple time varying signal paths between mobile users and the base station, which increase the reliability of communication. For an example, the multiple paths can be utilized to transmit multiple redundant copies of the data stream (space time codes) to the receiver to ensure reliable decoding [89]. This is called the *diversity gain*. Alternately, these multiple paths can be utilized to transmit different data streams from the same radio resource in separate spatial dimensions, which yields a *degree-of-freedom gain* (alternatively called multiplexing gain). Analysis of

this multiplexing gain started with the *single user MIMO* scenario, where the base station communicates with a single user through multiple transmit antennas. If there are N_t antennas at the base station and N_r antennas at the mobile, the base station could obtain a multiplexing gain of $\min(N_t, N_r)$ for this scenario [26]. This gain is constrained by the size and cost limitation of a user terminal which requires the number of receiver antennas to be minimum. Therefore, research focus shifted to *multi-user MIMO* which allows the base station to obtain the maximum multiplexing gain through transmitting to multiple users simultaneously [22,30]. This model is also called the *MIMO Gaussian broadcast channel*.

Some of the gains associated with MIMO can be harvested without requiring any knowledge of the wireless channel states, but some others can be exploited effectively only through some form of channel state information (CSI) at the base station [48, 85]. Consider a multiuser communication system utilizing multiple transmission antennas at the base station in a fading environment. The channel changes over time, and the goal of the base station is to maximize downlink data rates by taking channel variations into account. In this setting, selecting the users with the best instantaneous channel for communication is a simple communication strategy that is heuristically expected to maximize downlink data rates. Indeed, this is the classical communication approach utilizing multiuser diversity to take advantage of changing channel conditions for rate maximization, e.g., see [42] where a power control scheme for maximizing the uplink capacity is presented thanks to multiuser diversity. In such communication instances necessitating the use of CSI for adaptive signaling, feedback is an important means to convey required information from mobile users to the base station.

The capacity region of a Gaussian broadcast channel, which utilizes the aspects of multiuser diversity, is well studied in the literature [11,94,96,98]. It is shown in [98] that if the base station has full CSI, the capacity region of the Gaussian MIMO broadcast channel can be obtained through a technique called dirty paper coding (DPC) [16]. In DPC, the base station calculates the interference at each user, and transmits in a manner which minimizes this interference. Implementation of such a technique for MIMO systems is extremely complex computationally, which makes it infeasible for any practical purpose.

Therefore, most of the research in the area focused on linear beamforming techniques [60, 63, 86, 106] to achieve higher rates of data communication, which are provably near the achievable capacity limits through DPC.

1.2 Why Opportunistic Beamforming ?

Although these beamforming techniques reduce the complexity associated with DPC, they still require full CSI. This means, the base station should know all the vector channel gains between the transmit antennas and the users, and assuming N_t transmit antennas at the base station and a single antenna at each user, a user is required to feed back N_t complex values (or a quantized version of it [39]) to the base station. Of course, these techniques will work well in systems with a small number of users because obtaining the knowledge about all the vector channel gains will not be a huge burden on the feedback link. However, feasibility of obtaining this knowledge at the base station decreases with the number of users in the system, or in other words, the practicality of implementation decreases with the size of the system. This motivated schemes which operated on partial CSI.

Opportunistic beamforming is an extension of the conventional beamforming mechanism where full CSI is not required. It has attracted considerable attention starting with [97] since it is considered to be a practical way of reducing feedback requirements in MIMO systems, but still achieving the same performance limits with the DPC technique [79]. The capacity scaling laws achieved by opportunistic beamforming were first obtained by Shariff and Hassibi in [79]. Among many other results, they, in particular, show that if an opportunistic scheduling algorithm is used to harvest multiuser diversity gains, the downlink throughput scales optimally like $N_t \log\log n$, where n is the number of users in the system. To achieve these gains, their proposed scheduling algorithm operates as follows. The beams are randomly generated. Each mobile user feeds back the signal-to-interference-plus-noise-ratio (SINR) (*i.e.*, a real number) corresponding to the beam with the best signal quality and the index of this beam (*i.e.*, an integer). Upon receiving SINR values from all the users, the base station schedules the user having the

best SINR on each beam. This leads to a considerable reduction in feedback requirements when compared to signal processing techniques requiring full CSI to achieve system capacity.

The decrease in the feedback load will simultaneously decrease the latency of obtaining feedback. Most of the work on beamforming techniques assume that the channel does not change in the coherence interval. Therefore, accuracy of the results depends on the length of this coherence interval. In a practical setup, increased latency in the feedback loop may lead to the base station making user selection based on expired CSI, which will nullify any multi-user diversity gains.

On top of this, opportunistic beamforming has a low implementation complexity compared to signal processing techniques which require knowledge about the channel vectors between the base station and the users. Specifically, it will not require matrix inversions. The low complexity makes it easy to be implemented practically, even for a system with a large number of users.

Because of these reasons, although being introduced in 2002, there is still considerable research interest on this technique [6, 17, 40, 57, 62, 84, 104], and we will use it as a framework for our work. Interested readers are referred to [97] and [91] for more advantages of this technique.

1.3 Focus of the Thesis

In this thesis, we consider opportunistic communication along multiple orthonormal beams. The focus is on the total ergodic downlink communication rate, and the base station is provided only with partial CSI. The base station selects the user with the highest SINR on each beam to maximize the sum-rate at the downlink.

Although opportunistic beamforming reduces feedback considerably in comparison to having full CSI, the technique in its plain implementations still displays a waste of feedback resources. Consider a simple single beam scenario. In the feedback stage, all the users will feed back the SINR value on this beam to the base station. However, the base station will schedule communication only to the user with the highest SINR. Therefore,

there is a clear waste of resources created by the feedback of users having no realistic chance of being scheduled for communication.

Since all the users are required to feed back, there is a linear growth in feedback with the total number of users in the system. Therefore, the modest doubly logarithmic multiuser diversity gain at the downlink is achieved at the expense a $O(n)$ growth in feedback. The linear growth can again be an impractical burden on the uplink for a large numbers of users.

These considerations motivate the use of selective feedback techniques in which only the users with good channels are allowed to feed back [29]. By using selective feedback, we focus on a more stringent but practical finite limit on the feedback load. We assume that there is no cooperation between the different mobile users. Therefore, we only focus on the class of decentralized selective feedback policies. In this thesis, we ask: What is the structure of the sum-rate maximizing decentralized selective feedback policies, and how does the resulting sum-rate compare to the sum-rate without any user selection?

1.4 Contributions and the Outline of the Thesis

Firstly, we discuss some fundamentals and background information in Chapter 2. Obtaining the sum-rate maximizing decentralized selective feedback policy under finite feedback constraints provides an answer to the first part of the question raised in the previous section. We start the analysis in Chapter 3, where we set a finite constraint on the average number of users feeding back. Therefore, our initial optimization problem is finding the decentralized selective feedback policy which maximizes the ergodic downlink sum-rate, such that the average number of users feeding back is less than a given constant λ. This problem is not easy to solve since the optimization is done over function spaces [49], and the objective function is not necessarily convex.

We first show that any sum-rate maximizing selective decentralized feedback policy for a given constraint on the average number of users feeding back must be a *threshold feedback policy*. According to a threshold feedback policy, each user, independently from others, decides to feed back or not by comparing its SINR values with a predetermined

threshold value. Users are allowed to have different thresholds if such heterogeneity maximizes the total downlink rate. This result is highly adaptable due to its distribution independent nature, *i.e.*, it does not depend on the particular statistical model of the wireless channel as long as the resulting SINR distribution is continuous, which is true for most common fading models. It provides an analytical justification for the use of threshold feedback policies in practical systems, and reinforces previous work analyzing threshold feedback policies as a selective feedback technique without proving its optimality. It is robust to selfish unilateral deviations, and possesses a *stability* property from a game theoretic point of view. Finally, it reduces the search for rate-wise optimal feedback policies subject to the given feedback constraints from function spaces to a finite dimensional Euclidean space.

The above result forms a basis for the optimal threshold selection problem analyzed in Chapter 4. Now, the optimization problem is over the familiar finite dimensional Euclidean spaces. Each user will have its own threshold value. Now, the optimization problem is obtaining the set of threshold values which maximizes the ergodic downlink throughput, such that the average number of users feeding back is less than λ. Still, the optimization problem is a hard one to solve, due to the objective function not being necessarily convex as a function of the threshold values. Thus, we resort to the theory of majorization [51], and solve the optimal threshold selection problem by identifying an underlying *Schur-concave* structure in the sum-rate function. In particular, we obtain sufficient conditions for the Schur-concavity of the sum-rate, and therefore for the rate optimality of *homogenous* threshold feedback policies in which all users use the same threshold for their feedback decisions.

Rather surprisingly, our results reveal that the naive but intuitive approach of setting a homogenous threshold value to maximize the total downlink communication rate for a network with identical users experiencing statistically the same channel conditions does not always work. We provide a simple counterexample, where Schur-concavity of the sum-rate is violated, and it becomes strictly suboptimal to use the same threshold value to mediate the feedback decisions. An extensive numerical study utilizing our sufficient conditions is also performed to illustrate optimality and sub-optimality regions for the

homogenous threshold feedback policies for fading models other than Rayleigh fading such as Rician and Nakagami fading.

However, on the more positive side, we show that the sum-rate is always a Schur-concave function when two or more orthonormal beams are used to simultaneously communicate with multiple users located in a Rayleigh fading environment. In order to find answers to the second part of the question posed in the previous section, the difference between communication rates achieved with and without user selection is also illustrated for this fading scenario. In particular, when the threshold values are optimally set for large user populations, there is almost no rate loss even when the average number of users feeding back per beam is relatively small. From a practical point of view, this signifies a significant reduction in the feedback load without noticeable performance loss.

The results in Chapter 3 and Chapter 4 form a basis for the problem analyzed in Chapter 5. Now, we have obtained the optimality of a homogenous threshold value for a Rayleigh fading channel, where multiple orthonormal beams are transmitted from the base station. In Chapter 5, we show how to set this homogenous threshold value to ensure $O(1)$ feedback load on the average as the number of users in the system grows large. We illustrate the sub-optimal and optimal throughput scaling of such a system through the feedback-capacity tradeoff curve tracing the Pareto optimal boundary between feasible and infeasible feedback-capacity pairs. A feedback-capacity pair (λ, c) lying above this curve is an infeasible operating point, and a point below this curve is a feasible but suboptimal operating point. It is suboptimal in the sense that there is another feedback policy achieving the same capacity scaling with strictly less feedback. Our derived homogeneous threshold feedback policy achieves any point on this tradeoff curve. In particular, we show that there exists a sequence of homogeneous threshold feedback policies with appropriately chosen thresholds achieving any point on this curve. We also show that $N_t \log\log n$ scaling in [79] can be achieved by only allowing $O\left((\log n)^{\epsilon}\right)$ users to feed back on the average for any $\epsilon \in (0,1)$.

This feedback reduction is for a system with statistically identical users. In Chapter 5, we go on to study a system where different users in the cell experience different path-

loss gains. Obviously, the same threshold level for all users will not work for such a system because this will make the users located close to the base station feed back with very high probabilities compared to the users at the cell edge. In this chapter, we show that the threshold levels at different users can be systematically altered according to the user locations without violating the $O(1)$ feedback constraint. We obtain the feedback-capacity tradeoff curve for this system, and we again show to which extent the feedback requirement should be relaxed to achieve $N_t \log\log n$ scaling.

In Chapter 5, we also discuss how fairness can be achieved among users, which becomes an important issue for heterogeneous communication environments. If a beam is allocated to the user having the best SINR, a user situated far from the base station may starve for data compared to a user staying close to the base station. To address this issue, we introduce two new scheduling policies. These scheduling policies coupled up with the systematic alteration of the threshold levels at different users allow the system to achieve the optimum throughput scaling with the added advantage of ensuring fairness among users.

Most of our results in Chapter 5 are asymptotic in the sense that they are derived by letting the number of users in the system grow large. Hence, we also perform some numerical evaluations to illustrate the accuracy of the results for finite size systems. In particular, we show that the threshold levels set by using our asymptotic formulas clearly achieve $O(1)$ feedback in finite size systems as well. We also compare the two proposed fair scheduling policies with the scheduling policy of allocating the beam to the best user for finite n. Although achieving the same asymptotic performance, we observe that ensuring fairness causes a clear degradation in rate for any finite size system, which is, in fact, the tradeoff between rate maximization and ensuring fairness in a wireless communication network.

In Chapter 6, we make a subtle but interesting change to the optimization problem studied in Chapters 3, 4 and 5. In this chapter, we derive the structure of the optimal homogenous threshold feedback policy that maximizes the ergodic downlink sum-rate under a *peak feedback load constraint* λ, *i.e.*, an instantaneous constraint on the number of users feeding back. Firstly, we define a multiple access model which imposes this instan-

taneous feedback constraint on the system. Unlike Chapters 3, 4 and 5, where an ideal medium-access-control (MAC) layer for contention resolution on the uplink feedback channel is assumed and the likely packet collisions are neglected, we use a multi-packet reception model to resolve collisions from users in this chapter. The base station can reconstruct all the feedback packets successfully if and only if the random number of users feeding back is less than or equal to λ, which is the maximum number of packets that the base station can decode concurrently. We say that a collision occurs if the random number of users feeding back is greater than λ. In this case, all packets are destroyed together. We do not consider back-off and retransmission issues. Therefore, a collision will cause zero rate.

We show that this is a quasi-convex optimization problem by analyzing the rate expression, and solve it to obtain a formula for the optimal threshold value at the user. These results hold for most practical fading distributions as well. To discuss the results in more detail and to provide further insights, we apply them to the well known Rayleigh fading channel model. The discussion includes the implications of the derived formulas in practical systems. As we did in Chapter 4, we demonstrate the amount of feedback reduction that can be achieved without any noticeable performance degradation in rate by setting the threshold levels optimally, and as we did in Chapter 5, we illustrate the tradeoff between feedback and rate by obtaining the Pareto optimal boundary between feasible and infeasible feedback-rate pairs, but with an instantaneous constraint on the number of users feeding back.

We also study the rate-wise optimality of using a homogenous threshold level in Chapter 6. Again, in line with the justifications done in Chapter 4, a homogenous threshold level seems optimal intuitively because the users are statistically identical. However, we provide a simple counter example which shows that this intuition is not always true for an instantaneous feedback constraint as well. For some channel conditions, a better sum-rate can be achieved by setting threshold levels unequally among the users. Chapter 7 concludes the thesis.

1.5 Publications List

1.5.1 Journal Papers

1. T.N. Samarasinghe, H. Inaltekin, and J.S. Evans, "The Feedback-Capacity Tradeoff for Opportunistic Beamforming under Optimal User Selection," *Performance Evaluation*, 2012 .
2. T.N. Samarasinghe, H. Inaltekin, and J.S. Evans, "Optimal Selective Feedback Policies for Opportunistic Beamforming," *Submitted to IEEE Transactions on Information Theory*, July 2011 (revised August 2012).

1.5.2 Conference Papers

1. T.N. Samarasinghe, H. Inaltekin, and J.S. Evans, "Optimal Selective Feedback Policies for Opportunistic Beamforming Under Peak Feedback Load Constraints," in *Proc. IEEE International Symposium on Information Theory*, pp. 2919-2923, Boston, USA, July 2012.
2. T.N. Samarasinghe, H. Inaltekin, and J.S. Evans, "Vector Broadcast Channels: Optimal threshold selection problem," in *Proc. IEEE International Symposium on Information Theory*, pp. 1906-1910, Saint Petersburg, Russia, July-August 2011.
3. T.N. Samarasinghe, H. Inaltekin, and J.S. Evans, "Vector Broadcast Channels: Optimality of Threshold Feedback Policies," in *Proc. IEEE International Symposium on Information Theory*, pp. 1292-1296, Saint Petersburg, Russia, July-August 2011.
4. T.N. Samarasinghe, H. Inaltekin, and J.S. Evans, "The Feedback-Capacity Tradeoff for Opportunistic Beamforming," in *Proc. International Conference on Communications*, pp. 1-6, Kyoto, Japan, June 2011.
5. H. Inaltekin, T.N. Samarasinghe, and J.S. Evans, "Rate Optimal Limited Feedback Policies for the MIMO Downlink," in *Proc. International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, pp. 395-400, Princeton, USA, May 2011.

Chapter 2
Fundamentals and Background

This chapter will first explain the importance of channel state information (CSI) in the context of a wireless communication network, and the possible gains that can be achieved by having CSI at the base station. Then, we will discuss the difficulties associated with obtaining full CSI at the base station, which motivates the concept of opportunistic beamforming. We will explain the basic operation of opportunistic beamforming by defining a formal multi-beam single cell system model. This system model will be used through out this thesis as a framework for the analysis. We obtain an expression for the ergodic downlink throughput, and do a preliminary analysis of it for a Rayleigh fading channel model. We end the chapter with a detailed review on opportunistic beamforming and other related literature.

2.1 The Importance of Channel State Information (CSI)

IN a wireless communication system comprised of a multitude of mobile users (MUs), the base station being able to select and communicate at a higher rate with a user having a good channel state compared to a user having a poor one is referred to as the multiuser diversity gain [42]. This can be best illustrated through a simple example using the broadcast channel given in Fig. 2.1. The base station has one transmit antenna, and there are n MUs with a single receive antenna. Consider all the variables to be real numbers for simplicity. If we assume that the transmit power is one, the received signal by MU k is given by

$$Y_k = h_k s + Z_k,$$

where s is the transmitted symbol, h_k is the random channel gain between the base station and MU k, and Z_k is the additive white Gaussian noise with a variance of σ^2 at MU k. We

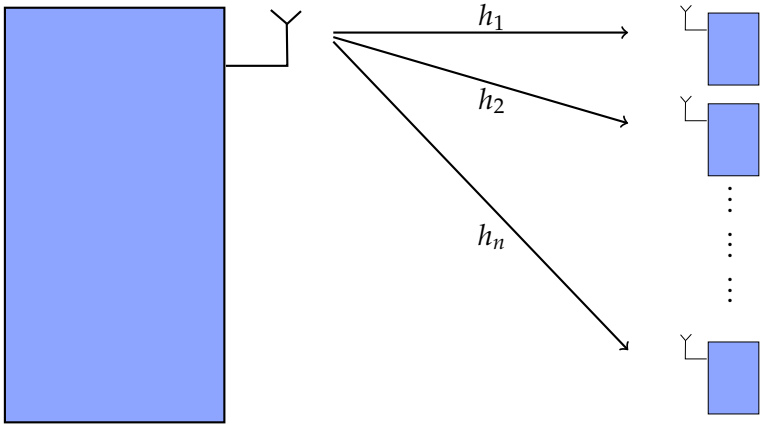


Figure 2.1: Scalar broadcast channel.

have dropped the time index for simplicity. Since the transmitted signal is a scalar, this channel model is well known as the scalar broadcast channel model.

If the base station knows all the channel gains between the base station and the MUs, *i.e.*, it knows h_k for all $k \in \{1, \ldots, n\}$, the MU with the best channel state can be picked for data communication to maximize the downlink rate. This maximum ergodic rate can be written as

$$R = \mathsf{E}\left[\log\left(1 + \max_{1 \le k \le n} \frac{|h_k|^2}{\sigma^2}\right)\right], \tag{2.1}$$

where the expectation is over the random channel gains, and we have assumed $\mathsf{E}\left[s^2\right] = 1$. However, to achieve this rate, the base station needs to know the channel gains of all the individual MUs.

Next, let us consider the scenario where the base station has no knowledge about the channel, *i.e.*, no CSI. The base station can now select a MU randomly for communication, and the maximum ergodic rate can be written as

$$R = \mathsf{E}\left[\log\left(1 + \frac{|h_k|^2}{\sigma^2}\right)\right]. \tag{2.2}$$

The rate achieved for each of these scenarios for different number of MUs are illustrated in Fig. 2.2. The figure shows clearly that having CSI is extremely useful since not having any CSI will nullify the multiuser diversity gain, and the system will achieve the same

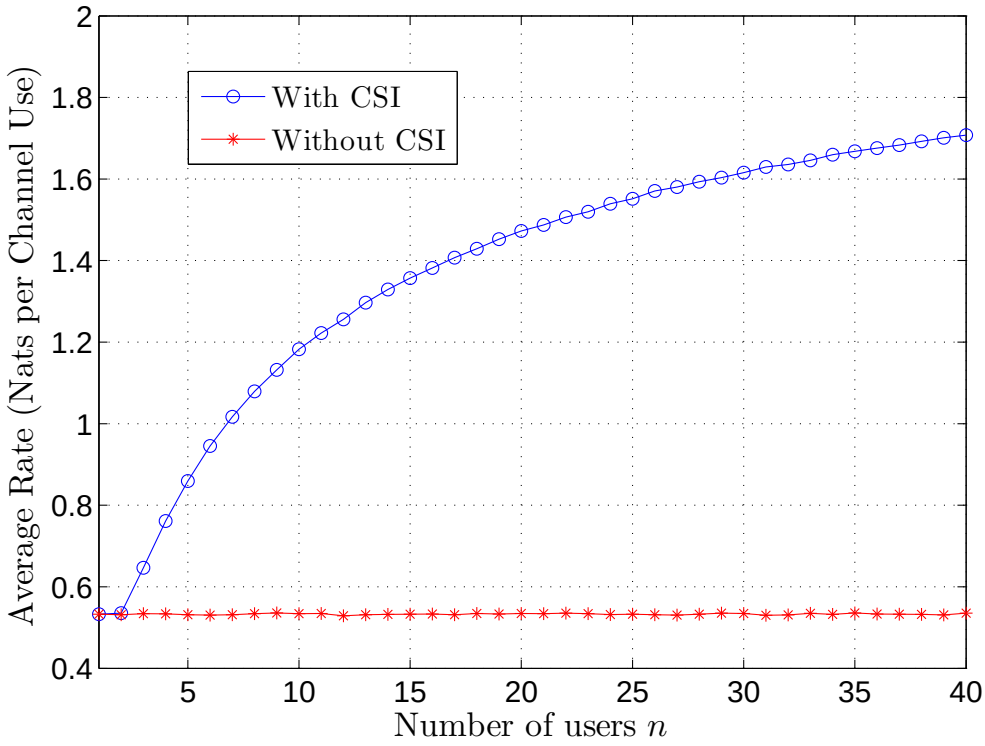


Figure 2.2: The rate achieved with and without CSI for different n.

ergodic rate irrespective of the number of MUs. this shows that CSI plays an important role in a wireless network.

Now, let us take a step further and consider a scenario where the base station has multiple transmit antennas, and each MU has a single receive antenna. The base station utilizes the multiple antennas for beamforming, and transmits a single beam. The channel model is illustrated in Fig. 2.3. This is called a vector broadcast channel since the transmitted signal is now a vector, defined by

$$\boldsymbol{x} = \boldsymbol{b}s,$$

where s is the transmitted symbol, $\boldsymbol{b}$ is called the beamforming vector, and we have again assumed that the transmit power is one for simplicity. Vector $\boldsymbol{b} = [b_{1,1}, b_{2,1}]^\top$ contains two complex numbers with which s is multiplied at each of the antennas (refer Fig. 2.3) to create the beam, thus called the beamforming vector. The received signal at MU k can

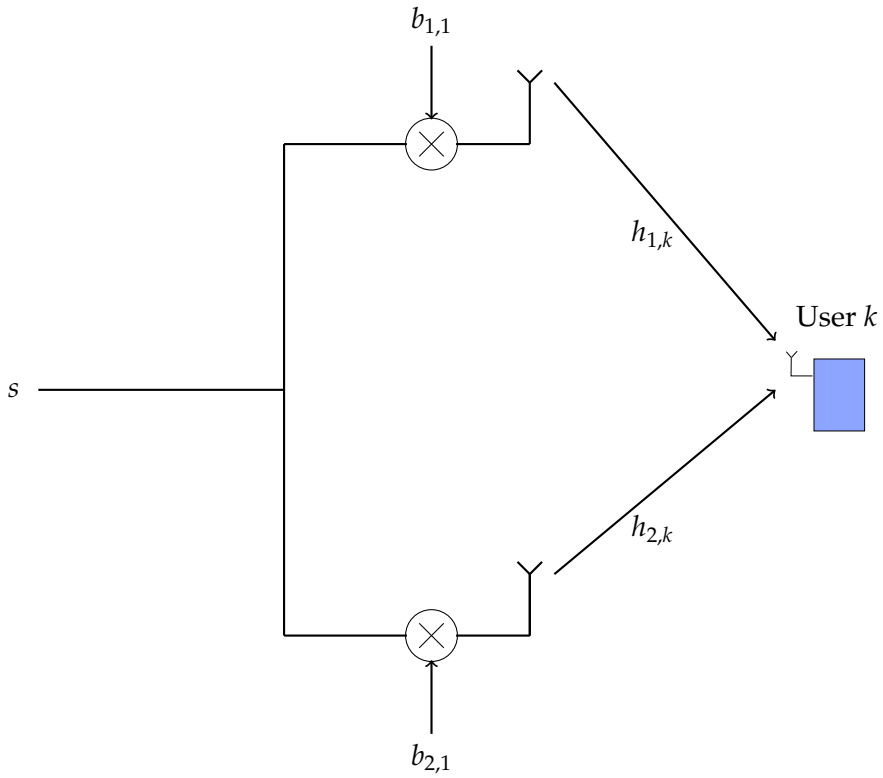


Figure 2.3: Vector broadcast channel: single beam and two transmit antennas.

be written as

$$Y_k = \boldsymbol{h}_k^\top \boldsymbol{b} s + Z_k,$$

where $\boldsymbol{h_k}$ is the random complex channel gain vector between the base station and the kth MU given by $\boldsymbol{h}_k = [h_{1k}, h_{2k}]^\top$, and Z_k is the complex white Gaussian noise at MU k having a variance of σ^2.

Firstly, lets assume that $\boldsymbol{b}$ is fixed and the base station knows all the vector channel gains between the base station and the MUs. Now, the maximum ergodic rate for this configuration can be written as

$$R = \mathsf{E}\left[\log\left(1 + \max_{1\leq k\leq n} \frac{|\boldsymbol{h}_k^\top \boldsymbol{b}|^2}{\sigma^2}\right)\right], \tag{2.3}$$

where again the expectation is over the random channel gains, and we have assumed $\mathsf{E}\left[s^2\right] = 1$. Using the CSI, the base station communicates with the MU who maximizes

$|\boldsymbol{h}_k^\top \boldsymbol{b}|^2$.

Next, let us consider $\boldsymbol{b}$ to be no longer fixed. Now, we can optimize over $\boldsymbol{b}$ to further increase the downlink rate, and obviously, CSI can be used for this process as well. Among many methods, the simplest will be to set $\boldsymbol{b}$ such that it maximizes the dot product in (2.3) (maximizing the SNR), which is setting it equal to the complex conjugate of the channel of that particular user. For this scenario, *i.e.*, when setting

$$\boldsymbol{b} = \frac{\boldsymbol{h}_k^\dagger}{|\boldsymbol{h}_k|},$$

where the operator $(\cdot)^\dagger$ represents the complex conjugate, the received signal by MU k can be given by

$$Y_k = |\boldsymbol{h}_k| s + Z_k.$$

Now, the channel gains from each of the antennas add up constructively (in phase) at the receiver. The maximum ergodic rate for this configuration can be achieved by maximizing among the MUs, and it is given by

$$R = \mathsf{E}\left[\log\left(1 + \max_{1\leq k\leq n} \frac{|\boldsymbol{h}_k|^2}{\sigma^2}\right)\right]. \tag{2.4}$$

From this two simple examples, we have explained the importance and the advantages of having CSI at the base station. However, the next important question is how CSI can be acquired. In our work, we consider frequency division duplexing (FDD). Therefore, the channel gains on the uplink and the gains on the downlink will be different. In such a setting, each MU/receiver has to calculate the channel gain on the downlink, and feed it back. In a more general setting where N_t transmitting antennas are employed at the base station, each MU has to feedback N_t complex numbers for the base station to have full knowledge about the channel (full CSI). Therefore, although beamforming techniques reduce the complexity compared to DPC, the full CSI requirement creates a huge feedback requisite for a large network. We conclude full CSI may be impractical for a large system, but having no CSI will mitigate the multiuser diversity gains as shown in Fig, 2.2. This problem motivates systems which operate on partial CSI.

2.2 Opportunistic Beamforming

Opportunistic beamforming is a communication technique relying on partial CSI, and has attracted considerable attention and research effort since its inception by Viswanath et al. in [97]. In this section, we will discuss some fundamentals of opportunistic beamforming with some analysis on the downlink throughput.

The signal model studied in [97] is similar to the one studied for a vector broadcast channel in the previous section (please refer to Fig. 2.3 for a graphical representation). The difference is, now, the beamforming vector $\boldsymbol{b}$ is generated randomly without any knowledge of the channel. Each MU calculates the SNR on this randomly generated beam, where

$$\mathrm{SNR}_k = \frac{|\boldsymbol{h}_k^\top \boldsymbol{b}|^2}{\sigma^2}$$

is the SNR at MU k, and feeds it back. The base station schedules communication to the MU with the highest SNR. Therefore, the amount of feedback is reduced from $N_t \times n$ complex numbers to n real numbers. The maximum ergodic rate can be written as

$$R = \mathsf{E}\left[\log\left(1 + \max_{1\leq k\leq n} \mathrm{SNR}_k\right)\right]. \tag{2.5}$$

Although reducing feedback, this model does not fully utilize the degree-of-freedom gain of a MIMO broadcast channel. To show this, we provide another simple numerical example, which is presented in Fig. 2.4. In this example, we have considered a system consisting of 4 transmit antennas, and simulated the ergodic rate for different number of beams M. As shown in Fig. 2.4 the rate can be significantly increased by increasing the number of simultaneous beams, *i.e.*, using the degree-of-freedom gain of the MIMO broadcast channel to communicate with multiple MUs simultaneously (multiuser MIMO). By using this concept, [79] extended opportunistic beamforming to simultaneous multiple orthogonal beams from the base station. They called this technique random beamforming. Since the multiple beams create interference, each MU feeds back the signal-to-noise-plus-interference-ratio (SINR) and the index of the best beam to the base station. Still, compared to having full CSI, the amount of feedback is reduced from $N_t \times n$ complex numbers to n real numbers and n integers, still providing considerable

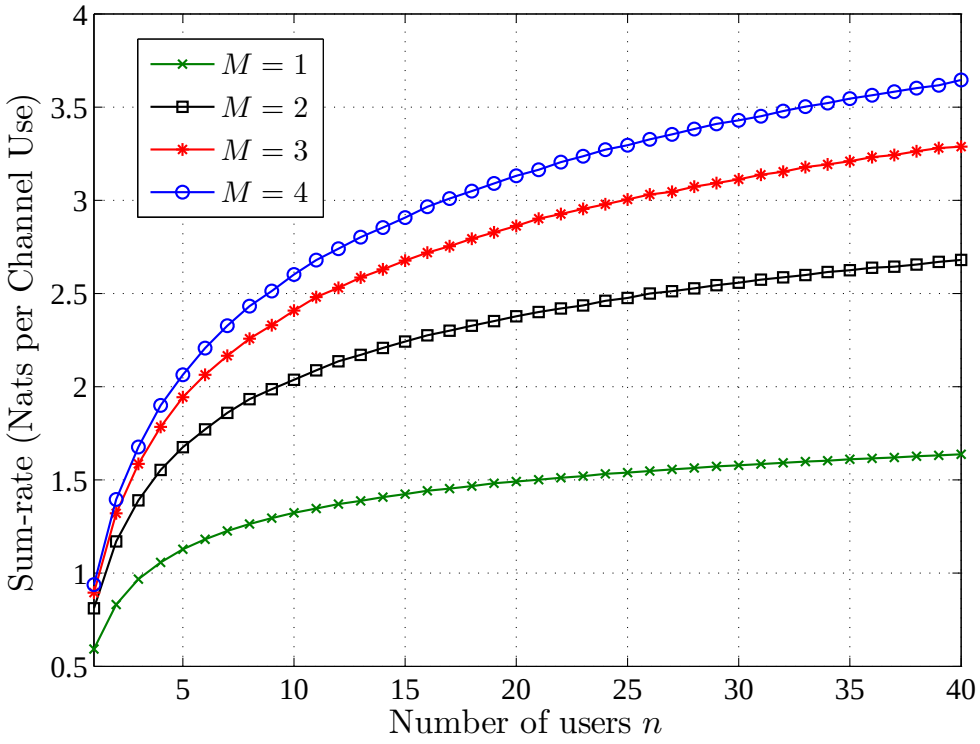


Figure 2.4: The behavior of the sum-rate with n for different values of M, where $N_t = 4$.

gains in feedback. In the next subsection, a formal system model is defined for such a system transmitting multiple orthogonal beams.

2.2.1 System Model

Consider a multi-antenna single cell vector broadcast channel. There are n MUs in the cell. The base station has N_t transmit antennas, and each MU is equipped with a single receive antenna. The channel gains between the receive antenna of the ith MU and the transmit antennas of the base station are given by

$$\boldsymbol{h}_i = (h_{1,i}, \ldots, h_{N_t,i})^\top, \tag{2.6}$$

where $h_{k,i}$ is the channel gain between the kth transmit antenna at the base station and the receive antenna at the ith MU. The channel gains $h_{k,i}$, $k = 1, \ldots, N_t$ and $i = 1, \ldots, n$, are

independent and identically distributed (i.i.d.) random variables. In addition, a quasi-static block fading model in which channel gains are constant through a coherence time interval, and change from one coherence period to another independently according to a common fading distribution is studied. For the sake of notational simplicity, the time index is dropped here in the channel model, and also later in the representation of transmitted and received signals.

The base station transmits M, $M \leq N_t$, different data streams intended for M different MUs. The symbols of the kth stream are represented by s_k. They are chosen from the capacity achieving unit power (complex) Gaussian codebooks, and are sent along the directions of M orthonormal beamforming vectors

$$\left\{ \boldsymbol{b}_k = (b_{1,k}, \ldots, b_{N_t,k})^\top \right\}_{k=1}^{M}. \tag{2.7}$$

These beamforming vectors can be either deterministic, or randomly generated and updated periodically. The overall transmitted signal from the base station is given by

$$\boldsymbol{s} = \sqrt{\rho} \sum_{k=1}^{M} \boldsymbol{b}_k s_k, \tag{2.8}$$

where ρ is the transmit power per beam [1]. The signal received by the ith MU is equal to

$$Y_i = \sqrt{\rho} \sum_{k=1}^{M} \boldsymbol{h}_i^\top \boldsymbol{b}_k s_k + Z_i, \tag{2.9}$$

where Z_i is the unit power (complex) Gaussian background noise. With these normalized parameter selections, ρ also signifies the SNR per beam as in [79]. Let $\gamma_{m,i}$ be the SINR value corresponding to the mth beam at the ith MU. Then, it is given by

$$\gamma_{m,i} = \frac{|\boldsymbol{h}_i^\top \boldsymbol{b}_m|^2}{\rho^{-1} + \sum_{k=1,k\neq m}^{M} |\boldsymbol{h}_i^\top \boldsymbol{b}_k|^2}. \tag{2.10}$$

Let $\boldsymbol{\gamma}_i = (\gamma_{1,i}, \ldots, \gamma_{M,i})^\top \in \mathbb{R}_+^M$ represent the SINR vector at MU i. Beams are sta-

[1] The transmit signal can be normalized by the number of beams to avoid the linear increase in transmit power with the number of beams. This normalization will not affect our results, so avoided for simplicity.

tistically identical, and the elements of γ_i are identically distributed for all $i \in \{1, \ldots, n\}$ with a common marginal distribution F. However, SINR values at a particular MU are dependent random variables.

2.2.2 Obtaining an Expression for Rate

A naive approach of allocating beams is randomly assigning them among the MUs. Then, the ergodic downlink sum-rate is given by

$$R = \mathsf{E}\left[\sum_{m=1}^{M} \log\left(1 + \gamma_{m,i}\right)\right], \tag{2.11}$$

where the expectation is taken over the random SINR values. Since the beams are statistically identical, we have

$$R \leq M \log\left(1 + \mathsf{E}\left[\gamma_{m,i}\right]\right), \tag{2.12}$$

using the Jensen's inequality. From (2.10), it is not hard to see that the SINRs behave as

$$\gamma_{m,i} \approx \frac{1}{M-1}$$

on the average for large M, by using the law of large numbers. Therefore, if we randomly assign beams to the MUs, the average downlink throughput R will be less than one,

$$\begin{aligned} R &\leq M \log\left(1 + \mathsf{E}\left[\gamma_{m,i}\right]\right) && (2.13)\\ &\approx M \log\left(1 + \frac{1}{M-1}\right) && (2.14)\\ &< \frac{M}{M-1} \approx 1, && (2.15) \end{aligned}$$

which means that we will not get a multiplexing gain even though M different signals are transmitted simultaneously. This is another example which clarifies that no CSI at the transmitter will be almost useless.

Because of this reason, the MUs feed back their maximum SINR value and the index

of the best beam. Each beam is allocated to the MU having the best SINR on it. Let $\mathcal{G}_m$ be the random set of MUs feeding back requesting beam m. Then, the ergodic downlink sum-rate can be written as

$$R = \mathsf{E}\left[\sum_{m=1}^{M} \log\left(1 + \max_{i\in\mathcal{G}_m} \gamma_{m,i}\right)\right]. \tag{2.16}$$

An Important Simplification of the Expression for Rate

We can do a simplification to this expression by assuming that the same MU will not have the maximum SINR value on two or more beams. Then the ergodic downlink sum-rate can be written as

$$R_{up} = \mathsf{E}\left[\sum_{m=1}^{M} \log\left(1 + \max_{1\le i\le n} \gamma_{m,i}\right)\right]. \tag{2.17}$$

In fact, this is an upper bound to the sum-rate achieved in (2.16). However, the assumption made will hold in all practical cases as discussed in Chapter 3 (Lemma 3.5). This assumption is very important because it makes the expression obtained for the rate mathematically tractable. R_{up} can be evaluated as

$$R_{up} = M \int_0^\infty \log(1+x) dG(x) dx, \tag{2.18}$$

where $G(x)$ is the cumulative distribution function (CDF) of the maximum of n i.i.d. random variables having a common marginal distribution F. $G(x)$ can be obtained by raising F to the power of n [68]. Therefore,

$$R_{up} = M \int_0^\infty \log(1+x) n F(x)^{(n-1)} dF(x). \tag{2.19}$$

Comparatively, evaluating R in 2.16 is not so straightforward. This is complex because in getting the CDF, we have to first consider the maximization of the SINR among the beams, which is the maximization of M dependent values, and then, the maximization among the MUs in $\mathcal{G}_m$, which is a random set. Because of this complexity, the upper bound R_{up} has been analyzed in many works associated with opportunistic beamform-

ing, after making the assumption that the same MU will not have the maximum SINR value on two or more beams [36, 69, 105] .

In the next subsection, we will study the throughput scaling of the ergodic sum-rate of opportunistic beamforming given in (2.16), for a Rayleigh fading channel.

2.2.3 Analysis on Rate: Rayleigh Fading Channels

We have already discussed the advantages of opportunistic beamforming in terms of feedback. However, the question is to what extent we should sacrifice in terms of rate to obtain this gain in feedback. It is shown in [70, 79] that the reduction in feedback doesn't hinder the performance of a system with a large number of MUs in terms of sum-rate capacity, and it achieves the same throughput scaling law as using dirty paper coding with full CSI at the transmitter [80]. To give a sketch of this proof, we will consider a Rayleigh fading environment, *i.e.*, $h_{k,i}, k = 1, \ldots, N_t$ and $i = 1, \ldots, n$, are assumed to be i.i.d. with the common distribution $\mathcal{CN}(0,1)$, where $\mathcal{CN}\left(\mu, \sigma^2\right)$ represents the *circularly-symmetric complex Gaussian* distribution with mean μ and variance σ^2.

For this channel model, the SINR distribution function F and the associated probability density function f can be given as

$$F(x) = 1 - \frac{e^{-\frac{x}{\rho}}}{(x+1)^{M-1}} \tag{2.20}$$

and

$$f(x) = \frac{e^{-\frac{x}{\rho}}}{(x+1)^{M}} \left[\frac{1}{\rho}(x+1) + M - 1\right], \tag{2.21}$$

respectively [79]. The proof techniques of these two expressions can be found in Chapter 5 (Appendix 5.9.1). For such a setup, a closed form expression for (2.19) is given in [105] and [36], and a closed form approximation can be found in [69].

As we have claimed earlier, the rate expression given in (2.17) is in fact an upper bound on the actual rate obtained using opportunistic beamforming. Now, we consider a feedback mechanism where a threshold is set at the receiver, and the MU feeds back

only if the SINR value is above the threshold. We set the threshold at one. Since we further reduce the amount of feedback setting the threshold, the rate obtained with the threshold is a lower bound to the rate given in (2.16). This lower bound can be written as

$$R \geq M\mathsf{Pr}\left\{\max_{1\leq i\leq n} \mathrm{SINR}_{i,1} \geq 1\right\} \mathsf{E}\left[\log\left(1+\max_{1\leq i\leq n} \mathrm{SINR}_{i,1}\right) \mid \max_{1\leq i\leq n} \mathrm{SINR}_{i,1} \geq 1\right],$$

which simplifies to

$$R \geq M\left[1-\left(\mathsf{Pr}\left\{\mathrm{SINR}_{i,1} \leq 1\right\}^n\right)\right] \mathsf{E}\left[\log\left(1+\max_{1\leq i\leq n} \mathrm{SINR}_{i,1}\right) \mid \max_{1\leq i\leq n} \mathrm{SINR}_{i,1} \geq 1\right],$$

since the MUs are i.i.d., and

$$\mathsf{Pr}\left\{\max_{1\leq i\leq n} \mathrm{SINR}_{i,1} \leq 1\right\} = \mathsf{Pr}\left\{(\mathrm{SINR}_{i,1} \leq 1)\right\}^n.$$

To prove the throughput scaling, [79] uses these bounds on the average rate, and show both the lower and upper bounds grow like $N_t \log\log n$, which is the same throughput scaling that one can achieve using dirty paper coding.

Please refer Fig. 2.4 to see how the rate behaves with the number of MUs in the system. The rate increases with the number of MUs because of the multiuser diversity gains, but when n is large, it increases slowly like $\log\log n$. A similar behavior is shown in Fig. 2.5, where the behavior of the rate with the number of beams is illustrated. The rate first increases with the number of beams because of degrees of freedom gain. However it later saturates because increasing the number of beams increases the interference, which simultaneously reduces the SINR values.

2.3 Literature Review

Some of the literature associated with opportunistic beamforming has already been introduced and discussed in the previous sections of this chapter. To summarize, the concept of opportunistic beamforming was introduced by Viswanath et. al. in 2002 [97]. They studied opportunistic beamforming for a single beam from the base station in a single

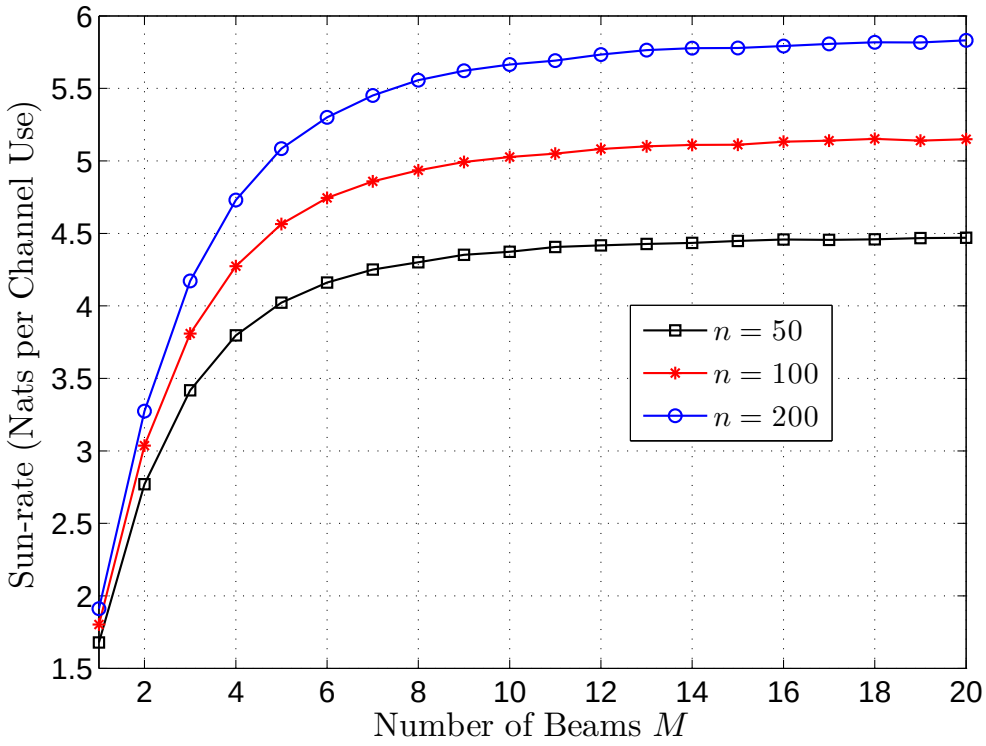


Figure 2.5: The behavior of the sum-rate with M for different values of n.

cell environment. Shariff and Hassibi extended this concept to simultaneous multiple orthogonal beams from a single base station, which they named *random beamforming* [79]. They showed that the proposed reduced feedback scheme can achieve the same capacity scaling laws of $N_t \log \log n$, with those achieved by dirty paper coding [80]. They also showed that the linear scaling of the sum-rate with the number of beams is preserved as long as N_t grows logarithmically with n.

Extensions and slight modifications of this result can be found in the literature as well. [101] gives a tighter upper bound on the average rate obtained in [79] using the properties of the extreme value distribution. In 2010, [70] analyzed the average rate of the exact scheme proposed in [79] without analyzing the bounds, and obtained the same result. They used order statistics [19] of the ratio of a linear combination of exponential random variables for this proof. The effect of spacial correlation between transmit antennas on this sum-rate capacity scaling has been studied in [2].

A system with multiple antennas at the receiver is studied in [79] as well. Let N_r be the number of receive antennas. The authors consider three different cases:

i) Treating each receiver antenna as an independent MU, *i.e.*, nN_r MUs.

ii) Assigning at most one beam to each MU, where each MU only feeds back information regarding the best receive antenna.

iii) Assigning N_r beams for the selected MUs assuming $\frac{M}{N_r}$ is an integer.

In the analysis the authors claim that case (i) is the best and case (iii) is the worst for finite n. However, all cases give similar results when n grows large asymptotically. The technique of considering each antenna as an independent MU is used in [20, 21] as well. This technique is improved in [70] to allow combining at the receivers in order to increase the received SINR. [70] uses the multiple receive antennas and a MMSE receiver, which is known to be the optimal linear receiver.

One of the main concerns associated with opportunistic beamforming is fairness, *i.e.*, whether a MU staying near the base station will be scheduled frequently and the MUs at the cell edge will be starved. To address this issue, an algorithm called *proportionally fair algorithm* is introduced in [97]. This takes into consideration the ratio between the requested rate and the average throughput for a particular MU. A MU with a high ratio will be prioritized over the others. If a MU is not scheduled for a long time the average throughput for the MU will reduce giving a higher value for the ratio, for that particular MU. In [79], the authors claim that fairness can be obtained automatically as a byproduct of having multiple beams. A MU located near the base station gets a high signal power as well as high interference from the other beams. The MUs at the cell edge gets low signal power and low interference. They claim that the fairness will prevail since the selection parameter is the SINR. However, this is not a fair assumption in a multi-cell environment. Therefore, in [54], the proportional fair algorithm has been extended for multiuser scheduling.

Another concern associated with opportunistic beamforming is its applicability to finite systems. Various methods have been proposed to optimize opportunistic beamforming for smaller networks as well. It has been shown in [56] and [55] that opportunistic beamforming can be effectively used for sparse networks. The main idea in these works

is to select a target group of MUs using opportunistic beamforming, and then use more efficient beamforming schemes to serve the MUs in the target group. The more efficient methods include transmit beamforming and also allocating power among beams optimally [87]. In [54], it has been shown how feedback aggregation can be used to optimize performance of opportunistic beamforming. In this scheme, the base station uses the memory about the channel to optimally select the beamforming vectors. The base station keeps track of beamforming vectors which give higher throughput and use them more frequently. In [15,41,64], schemes are proposed where multiple beamforming vectors are used to optimize the performance. In this technique, the time-slot is sub divided into Q mini-slots, with a separate set of beamforming vectors for each mini-slot. Rate is checked on each of the mini-slots and the optimum set of beamforming vectors is selected for transmission.[92] proposes adaptive beam selection, where scheduler picks the optimum subset of beams that maximizes the system sum-rate instead of communicating on all the generated beams.

Natural ways of further reducing the feedback load of a MIMO broadcast channel are using quantized feedback [28,35,39,43,90], or exploiting the properties of the fading process for feedback compression [24, 25, 38, 61]. With the developments of digital communication, quantized feedback is preferred in many communication systems. A comparison between digital and analog feedback methods is given [10]. Before long, quantized feedback techniques were introduced and studied using an opportunistic beamforming framework [21, 65, 78, 93]. Out of them, [65] investigate the performance of opportunistic beamforming with an SNR quantization scheme that maximizes the expected system throughput. They derive the SNR statistics and the system throughput to optimize the SNR quantization with respect to thresholds separating quantization levels. They show that when the number of feedback bits grow large, the PDF of the quantized SNR approximates well to the continuous (actual) SNR distribution. [78] extends the results in [43] for an opportunistic beamforming framework. They also propose and analyze a compound strategy that uses one bit for multi-user diversity and any further information for beamforming.

Although all of these techniques combined with opportunistic beamforming reduce

feedback considerably in comparison to having full CSI, still, all the MUs are required to feed back, which can be an impractical feedback requirement on the feedback channel. Also, it is a waste of feedback resources created by the MUs having no realistic chance of being scheduled for communication. [29] shows that most part of this feedback is unjustifiable, and selective feedback techniques, where only a fraction of the MUs feedback, can be used to achieve similar diversity gains through considerably lesser amount of feedback. [34, 45] extend the results in [29] by proposing a feedback mechanisms which operates on multiple feedback thresholds. By employing multiple feedback thresholds, the base station can conduct the feedback collection process by polling the MUs sequentially from the highest threshold value down to the lowest threshold value until feedback from one or more MUs is received. A threshold feedback policy was first introduced for opportunistic beamforming in [79]. A similar feedback policy was studied in [70]. According to this feedback policy, a MU feeds back if the SINR on the best beam is above a given constant threshold. However, even when using a constant threshold feedback policy, the number of MUs feeding back grows linearly with the total number of MUs in the system. As a result, to achieve modest doubly logarithmic multiuser diversity gains at the downlink through opportunistic scheduling, we still need to improve the capacity of the feedback link linearly with the number of MUs in the system. In [21], it was shown that it is enough to have only $O(\log n)$ MUs feeding back to achieve the same downlink sum-rate scaling in [79] by varying the common threshold level with the total number of MUs in the system.

With the emergence of wide-band communication, opportunistic beamforming has been extended for Orthogonal-frequency-division multiplexing (OFDM) networks as well [23, 27, 46, 66, 81]. For wide-band systems, the transmitted signals cannot form a single beam pointing to a certain direction due to the multipath propagation. Therefore, in OFDM networks, it is suggested to decompose the bandwidth into parallel narrow band sub-carriers so that random beamforming concept can be applied on each of them. [27] considers the neighboring sub-carriers as a cluster, and shows that the knowledge of the quality of the center subcarrier sheds light about the quality of the whole cluster. Therefore, they claim, MUs need only feedback the best SINR at the center subcarrier of each

cluster in order to maintain the same throughput scaling as when full CSI is available at the base station.

All the above works have been done in a single cell environment. Some analysis on opportunistic beamforming in multi cell environments exist as well. [4] and [88] have mainly concentrated on optimizing the scheduling strategies of opportunistic beamforming in a multi-cell environment. [59] focus on using combined scheduling and beamforming to mitigate performance losses due to inter-cell interference. [44, 58, 67] have concentrated on the throughput of a simple multi cell model with noncooperative base stations. [99] consider a multi cell random beamforming scheme where a cooperative base station can simultaneously beamform to its local MUs by utilizing full CSI and randomly beamform to the MUs in the neighboring cells.

More related works have been discussed in Chapters 3, 4, 5 and 6.

Chapter 3

Optimality of Threshold Feedback Policies

Beamforming techniques utilizing only partial channel state information (CSI) has gained popularity over other communication strategies requiring perfect CSI thanks to their lower feedback requirements. The amount of feedback in beamforming based communication systems can be further reduced through selective feedback techniques in which only the users with channels good enough are allowed to feed back by means of a decentralized feedback policy. In this chapter, we prove that thresholding at the receiver is the rate-wise optimal decentralized feedback policy for feedback limited systems with prescribed feedback constraints. This result is highly adaptable due to its distribution independent nature, provides an analytical justification for the use of threshold feedback policies in practical systems, and reinforces previous work analyzing threshold feedback policies as a selective feedback technique without proving its optimality. It is robust to selfish unilateral deviations. Finally, it reduces the search for rate-wise optimal feedback policies subject to feedback constraints from function spaces to a finite dimensional Euclidean space.

3.1 Introduction

We consider the classical opportunistic communication along multiple orthonormal beams. The focus is on the total downlink communication rate[1], and the base station is provided only with *partial* CSI (*i.e.*, downlink SINR values) for scheduling such as in the IS-856 standard. Hence, the (full CSI) sum-rate capacity achieving dirty paper precoding [11,16,94,96,98], or any other transmit beamforming strategy requiring full CSI to this end, is automatically disallowed. The wireless channels, and therefore

[1]Unless otherwise stated, we use the term *rate* (or, its derivatives such as sum-rate, total rate, aggregate communication rate) to always refer to the *ergodic rate* obtained by averaging over many fading states.

the attained signal-to-interference-plus-noise ratios (SINR) by different users on different beams, change over time. The base station selects the best user (the user with the highest SINR) per beam to maximize the sum-rate at the downlink.

This is the opportunistic beamforming approach utilizing multiuser diversity and varying channel conditions to extract all degrees-of-freedom available for the downlink communication (provided by the use of multiple transmit antennas) as well as to deliver improved power gains [79, 97]. Indeed, it achieves the same full CSI sum-rate capacity to a first order for large numbers of mobile users (MUs) in the network [79]. However, for large numbers of MUs, the opportunistic beamforming approach still requires large amounts of data to be fed back, which is an onerous requirement on the uplink feedback channel. What is needed is a selective decentralized feedback policy that will only choose a small subset of MUs to be multiplexed on the uplink feedback channel. In this case, the downlink sum-rate is certainly a function of the feedback policy selecting MUs. We ask: What is the structure of the sum-rate maximizing selective decentralized feedback policies, and how does the resulting sum-rate compare to the sum-rate without any user selection? This chapter focuses on selective feedback techniques, and formally establishes the structure of rate-wise optimal feedback policies for vector broadcast channels under finite feedback constraints.

3.1.1 Related Work

Feedback load reduction techniques for adaptive signaling in wireless communication networks have been a key area of research for more than a decade, especially with the advent of MIMO technology [18, 39, 43, 47, 48, 79]. Among many promising approaches proposed over the last decade, opportunistic beamforming has attracted considerable attention and research effort since its inception by Viswanath et al. in [97]. It is a practical way of reducing feedback requirements for vector broadcast channels, yet still achieves the full CSI sum-rate capacity at the downlink to a first order [79]. In this chapter, we are also motivated by such opportunistic communication and beamforming techniques, and focus on the downlink sum-rate maximization under finite feedback constraints on the feedback uplink.

Capacity scaling laws attained by opportunistic beamforming were first obtained by Shariff and Hassibi in [79]. Among many other results, they, in particular, showed that if an opportunistic scheduling algorithm is used to harvest multiuser diversity gains, the downlink throughput scales optimally like $N_t \log\log n$, where N_t is the number of transmit antennas at the base station, and n is the number of MUs in the system. In [101], the authors built upon [79] to derive tighter expressions for the downlink sum-rate scaling for opportunistic beamforming. Unlike these works, the results derived for the structure and optimization of the downlink sum-rate in this chapter are correct for any number (small and large) of MUs in the network. In addition, the sum-rate maximization problem addressed in this chapter does not appear in [79] and [101].

Without any user selection, the number of MUs feeding back grows linearly with the total number of MUs in the system to achieve double logarithmic growth in the downlink sum-rate. Threshold feedback policies are frequently used to alleviate such an excessive feedback requirement on the uplink [6, 21, 29, 34, 70, 73, 77, 79]. In [29], the authors proposed to use a common threshold level to arbitrate MUs' feedback decisions for scalar channels. They showed that this approach has the potential to significantly reduce the total feedback load on the uplink while maintaining *almost* the same sum-rate performance at the downlink. In [34], the authors extended the feedback scheme proposed in [29] by using multiple threshold levels. This work differs from [29] and [34] in three important aspects. Firstly, we provide an analytical justification for why threshold feedback policies are right choice for user selection, *e.g.*, see Section 3.3 for details. Secondly, in Chapter 4, we pose an optimum threshold selection problem in which we search for the optimum assignment of thresholds to MUs. We show that using the same threshold value for all MUs is not always optimum even if all MUs experience statistically the *same* channel conditions. Finally, our results are given for more general vector broadcast channels.

In [79] and [70], the authors used a constant threshold level, the same for all MUs and independent of the number of MUs, to reduce the total feedback load for vector broadcast channels within the opportunistic beamforming framework. Such a constant thresholding scheme cannot eliminate the linear growth in the average number of MUs feeding

back. In [21], it was shown that it is enough to have only $O(\log n)$ MUs feeding back to achieve the same downlink sum-rate scaling in [79] by varying the common threshold level with the total number of MUs in the system. The threshold was set such that any MU having a SINR above the threshold will achieve $\log\log n$ scaling, allowing one bit feedback [20, 21, 103]. This result was extended in [73, 77] by showing that $O\left((\log n)^{\epsilon}\right)$, $\epsilon \in (0, 1)$, MUs are enough to achieve the same downlink sum-rate scaling in [79]. It is *almost as if* constant feedback load is enough to maintain optimum sum-rate scaling but not exactly. These results will be discussed in more detail in Chapter 5. Recently, [6] studied the amount of feedback (partial CSI) required to achieve the same sum-rate capacity achieved with perfect CSI at the BS, considering various SNR regimes. In contrast to these previous works, we focus on more stringent but practical *constant* feedback requirements in this chapter. The sum-rate maximization framework introduced here does not exist in these papers, either. Finally, these previous works only focused on the asymptotic sum-rate scaling behavior, whereas our results are correct for any finite number of MUs in the network.

3.1.2 Contributions

In this chapter, we first show that any sum-rate maximizing selective decentralized feedback policy for a given constraint on the average number of MUs feeding back must be a *threshold feedback policy* in which each MU, independently from others, decides to feed back or not by comparing its SINR values with a predetermined threshold value. We start with a simple single user scenario, and then extend it to a general multi-user scenario, where different MUs are allowed to have different thresholds if such heterogeneity in thresholds maximizes the total downlink rate. This thresholding optimality result does not depend on the particular statistical model of the wireless channel as long as the resulting SINR distribution is continuous, which holds for most common fading models such as Rayleigh, Rician and Nakagami fading. It also possesses a *stability* property from a game theoretic point of view as explained in Section 3.5.

These findings provide an analytical justification for the use of threshold feedback policies in practical systems, and strengthen previous work on thresholding as an appro-

priate selective feedback scheme, *e.g.*, see [21, 29, 34, 70, 73, 79]. They also form a basis for the optimum threshold selection problem analyzed in Chapter 4. To some extent, our thresholding optimality result is intuitive and expected. It is even known to hold in the limit without feedback constraints for richly scattered Rayleigh fading environments [21, 73]. However, its proof in our case is not straightforward, and requires a careful analysis of rate gain and loss events due to coupling effects, induced by finite feedback constraints, of MUs' individual feedback rules on the sum-rate function. We show that the proof can be simplified considerably if we study a scenario where the MUs use a homogeneous feedback policy since the homogeneity nullifies the coupling effects of MUs' individual feedback rules on the sum-rate function.

The organization of this chapter is as follows. We precisely define feedback policies and formulate the problem of finding the optimal feedback policy maximizing aggregate communication rate under finite feedback constraints as a function optimization problem in Section 3.2. We prove that the rate-wise optimal feedback policy solving this optimization problem is a threshold feedback policy in Section 3.3. In Section 3.4, we analyze the special case, where the MUs are restricted to feed back using a homogenous threshold feedback policy. We discuss the results in Section 3.5, and Section 3.6 concludes the chapter.

3.2 Problem Formulation

We study the vector broadcast channel model given in Subsection 2.2.1. The base station communicates with n MUs through M different orthogonal beams simultaneously. The base station has N_t transmit antennas, and each MU is equipped with a single receive antenna. The beams are assumed to be statistically identical, and MUs experience statistically independent channel conditions. $\gamma_{m,i}$ is the SINR at beam m at MU i, and $\boldsymbol{\gamma}_i \in \mathbb{R}_+^M$ represents the SINR vector at MU i. The elements of $\boldsymbol{\gamma}_i$ are identically distributed for all $i \in \mathcal{N}$ with a common marginal distribution F, where $\mathcal{N} = \{1, \ldots, n\}$. We will assume that F is continuous, and has the density f with support $\mathbb{R}_+$, which are true for many fading models including Rayleigh, Rician and Nakagami fading. If $M = 1$, we will

use γ_i to denote the SINR of MU i on this single beam. The M-by-n SINR matrix of the whole n-user communication system is denoted by $\mathbf{\Gamma} \in \mathbb{R}_+^{M\times n}$. If the base station has perfect knowledge of $\mathbf{\Gamma}$, the aggregate communication rate can be maximized by choosing the best MU with the highest SINR on each beam. However, this necessitates excessive amount of feedback and information exchange between the base station and MUs. Therefore, we focus on the sum-rate maximization under finite feedback constraints, where MUs feed back according to a predefined selective feedback policy as defined below.

Let

$$\gamma_i^\star = \max_{1\le k\le M} \gamma_{k,i}$$

be the maximum SINR value at MU i, and

$$b_i^\star = \arg\max_{1\le k\le M} \gamma_{k,i}$$

be the index of the best beam achieving $\gamma_i^\star$. Also, let $\mathcal{M} = \{1,\dots,M\}$. Using these notations, we formally define a feedback policy as follows.

Definition 3.1. *A feedback policy* $\boldsymbol{\mathcal{F}} : \mathbb{R}_+^{M\times n} \mapsto \{\Omega \bigcup \{\text{ø}\}\}^n$ *is an* $\{\Omega \bigcup \{\text{ø}\}\}^n$*-valued function*

$$\boldsymbol{\mathcal{F}} = (\mathcal{F}_1,\dots,\mathcal{F}_n)^\top,$$

where $\mathcal{F}_i : \mathbb{R}_+^{M\times n} \mapsto \Omega \bigcup \{\text{ø}\}$ *is the feedback rule of MU* i*,* Ω *is the set of all feedback packets* [2] *and* ø *represents the no-feedback state. We call* $\boldsymbol{\mathcal{F}}$ *a* general decentralized feedback policy *if* $\mathcal{F}_i$ *is only a function of* $\boldsymbol{\gamma}_i$ *for all* $i \in \mathcal{N}$*. We call it a* homogenous general decentralized feedback policy *if it is decentralized and all MUs use the same feedback rule. Finally, we call it a* maximum SINR decentralized feedback policy*, if* $\mathcal{F}_i(\boldsymbol{\gamma}_i)$ *is only a function of* $\gamma_i^\star$ *and* $b_i^\star$*, and produces a feedback packet containing* $\gamma_i^\star$ *as the sole* SINR *information on a positive feedback decision, and otherwise produces* ø*.*

Intuitively, a feedback policy determines whether a MU will feed back or not. Upon a positive feedback decision, it generates a feedback packet containing SINR values at

[2] Many different feedback mechanisms can be considered. Feeding back all the SINR values, feeding back only the maximum SINR, feeding back the two largest SINRs, are just few of these approaches. The structure of the feedback packet will depend on the feedback approach. Ω is the set containing all of these possible feedback packets.

selected beams (along with other information to be contained in the packet header), and sends it to the base station for central processing. When it is clear from the context, we will omit the term "general". We will index system-wide feedback policies by superscripts such as $\boldsymbol{\mathcal{F}}^i$, and individual feedback rules by subscripts such as $\mathcal{F}_i$. We use the term "policy" to refer to system-wide feedback rules, whereas the term "rule" is used to refer to individual feedback rules. The definitions given for system-wide feedback policies extend to individual feedback rules in an obvious way when possible. We assume that there is no cooperation between different MUs, which is true for most practical systems, therefore we can narrow down our study to decentralized feedback policies for the system in consideration.

Furthermore, we will focus our attention on beam symmetric feedback policies since beams are assumed to be statistically identical. We formally define beam symmetric policies as follows.

Definition 3.2. *Let* $\Pi : \mathbb{R}^M \mapsto \mathbb{R}^M$ *be a permutation mapping,* i.e.,

$$\Pi(\boldsymbol{\gamma}) = \left(\gamma_{\pi(1)}, \ldots, \gamma_{\pi(M)}\right)^\top,$$

for some one-to-one $\pi : \mathcal{M} \mapsto \mathcal{M}$*. For* $\boldsymbol{\Gamma} \in \mathbb{R}^{M \times n}$*, let*

$$\Pi(\boldsymbol{\Gamma}) = \left[\Pi(\boldsymbol{\gamma_1}), \ldots, \Pi(\boldsymbol{\gamma_n})\right].$$

If $\mathcal{I}_i$ *is the set of beam indexes selected by* $\mathcal{F}_i(\boldsymbol{\Gamma})$*, and* $\pi(\mathcal{I}_i)$ *is the set of beam indexes selected by* $\mathcal{F}_i(\Pi(\boldsymbol{\Gamma}))$ *for all* $i \in \mathcal{N}$*, we say* $\boldsymbol{\mathcal{F}}$ *is a* beam symmetric *feedback policy.*

This symmetry assumption is just for the sake of notational simplicity, and the same techniques can be generalized to beam asymmetric policies by allowing different feedback policies for different beams at MUs. We let Ξ denote the set of all beam symmetric decentralized feedback policies. When it is clear from the context, we will also omit the term "beam symmetric".

Given a feedback policy $\boldsymbol{\mathcal{F}}$, we have a random set of MUs $\mathcal{G}_m(\boldsymbol{\mathcal{F}}(\boldsymbol{\Gamma}))$ requesting beam $m \in \mathcal{M}$. When $\mathcal{G}_m(\boldsymbol{\mathcal{F}}(\boldsymbol{\Gamma}))$ is a non-empty set at a given fading state, the base station selects the MU with the highest SINR in this set to maximize the instantaneous

communication rate in the direction of beam m. If $\mathcal{G}_m\left(\boldsymbol{\mathcal{F}}\left(\boldsymbol{\Gamma}\right)\right)$ is an empty set, we say a *feedback outage event* occurs at beam m, and zero rate is achieved at this beam.[3]

Then, the downlink ergodic sum-rate achieved under the feedback policy $\boldsymbol{\mathcal{F}}$ is given by

$$R\left(\boldsymbol{\mathcal{F}}\right)=\mathsf{E}_{\boldsymbol{\Gamma}}\left[r\left(\boldsymbol{\mathcal{F}},\boldsymbol{\Gamma}\right)\right]=\mathsf{E}_{\boldsymbol{\Gamma}}\left[\sum_{m=1}^{M}\log\left(1+\max_{i\in\mathcal{G}_m\left(\boldsymbol{\mathcal{F}}\left(\boldsymbol{\Gamma}\right)\right)}\gamma_{m,i}\right)\right], \tag{3.1}$$

where $r\left(\boldsymbol{\mathcal{F}},\boldsymbol{\Gamma}\right)$ is the *instantaneous* sum-rate achieved under the feedback policy $\boldsymbol{\mathcal{F}}$, expectation is taken over the random SINR matrices, and the result of the maximum operation is zero when $\mathcal{G}_m\left(\boldsymbol{\mathcal{F}}(\boldsymbol{\Gamma})\right)$ is an empty set. $r^m\left(\boldsymbol{\mathcal{F}},\boldsymbol{\Gamma}\right)$ and $R^m\left(\boldsymbol{\mathcal{F}}\right)$ denote the instantaneous sum-rate and the ergodic sum-rate on beam m, respectively. Note that

$$r^m\left(\boldsymbol{\mathcal{F}},\boldsymbol{\Gamma}\right)=\log\left(1+\max_{i\in\mathcal{G}_m\left(\boldsymbol{\mathcal{F}}\left(\boldsymbol{\Gamma}\right)\right)}\gamma_{m,i}\right), \tag{3.2}$$

and

$$R^m\left(\boldsymbol{\mathcal{F}}\right)=\mathsf{E}_{\boldsymbol{\Gamma}}\left[r^m\left(\boldsymbol{\mathcal{F}},\boldsymbol{\Gamma}\right)\right]. \tag{3.3}$$

Also, the sum-rate achieved on an event $\mathcal{A}$ under $\boldsymbol{\mathcal{F}}$ is written as

$$R\left(\boldsymbol{\mathcal{F}},\mathcal{A}\right)=\mathsf{E}_{\boldsymbol{\Gamma}}\left[r\left(\boldsymbol{\mathcal{F}},\boldsymbol{\Gamma}\right)\mathbf{1}_{\mathcal{A}}\right], \tag{3.4}$$

and conditioned on an event $\mathcal{A}$ (or, a random variable), we define the *conditional* sum-rate as

$$R\left(\boldsymbol{\mathcal{F}}|\mathcal{A}\right)=\mathsf{E}_{\boldsymbol{\Gamma}}\left[r\left(\boldsymbol{\mathcal{F}},\boldsymbol{\Gamma}\right)|\mathcal{A}\right]. \tag{3.5}$$

We will use $R\left(\boldsymbol{\mathcal{F}}\right)$ as the performance measure of a given feedback policy along the rate dimension.

[3]Note that the base station does not have access to any CSI on the feedback outage event. Without any CSI, reliable communication is still possible if we can average over very large time-scales for all MUs. The extra rate term to be added to (3.1) in this case would not affect our analysis in remainder of the chapter, and therefore is omitted for simplicity.

Given a feedback policy $\boldsymbol{\mathcal{F}}$, we will use the average number of MUs feeding back per beam $\Lambda(\boldsymbol{\mathcal{F}})$ to measure the performance of $\boldsymbol{\mathcal{F}}$ along the feedback dimension. $\Lambda(\boldsymbol{\mathcal{F}})$ can be written as

$$\Lambda(\boldsymbol{\mathcal{F}}) = \sum_{i=1}^{n} p_i, \tag{3.6}$$

where

$$p_i = \mathsf{Pr}\left\{\mathcal{F}_i(\boldsymbol{\Gamma}) \text{ selects beam } 1\right\} \tag{3.7}$$

since $\boldsymbol{\mathcal{F}}$ is beam symmetric. We are interested in maximizing the ergodic sum-rate under finite feedback constraints, and the resulting rate maximization problem can be written as

$$\begin{array}{ll} \underset{\boldsymbol{\mathcal{F}}\in\Xi}{\text{maximize}} & R(\boldsymbol{\mathcal{F}}) \\ \text{subject to} & \Lambda(\boldsymbol{\mathcal{F}}) \leq \lambda \end{array}, \tag{3.8}$$

i.e., find the optimal feedback policy maximizing the aggregate communication rate subject to feedback constraint λ. This optimization problem is over function spaces [49], and the objective function is not necessarily convex. In this chapter, we will reduce the search for optimal feedback policies to an optimal threshold selection problem over finite dimensional Euclidean spaces by proving rate-wise optimality of threshold feedback policies. The next section establishes the optimality of threshold feedback policies.

3.3 Optimality Analysis

In this section, we show that the solution of the optimization problem posed in (3.8) must be a threshold feedback policy. We start our analysis by formally defining threshold feedback policies.

Definition 3.3. *We say*

$$\boldsymbol{\mathcal{T}} = (\mathcal{T}_1, \ldots, \mathcal{T}_n)^\top$$

is a general threshold feedback policy *(GTFP) if, for all* $i \in \mathcal{N}$*, there is a threshold* τ_i *such that* $\mathcal{T}_i(\gamma_i)$ *generates a feedback packet containing* SINR *values* $\{\gamma_{k,i}\}_{k\in\mathcal{I}_i}$ *if and only if* $\gamma_{k,i} \geq \tau_i$ *for all* $k \in \mathcal{I}_i \subseteq \mathcal{M}$*. We call it a* homogenous general threshold feedback policy *if all MUs use the same threshold* τ, i.e., $\tau_i = \tau$ *for all* $i \in \mathcal{N}$.

We note that a MU can be allocated to multiple beams according to Definition 3.3. Another class of threshold feedback policies are the feedback policies limiting each MU to request only the beam with the highest SINR, *e.g.*, see [21,70,73,79]. We call this class of feedback policies maximum SINR threshold feedback policies, and formally define them as follows.

Definition 3.4. $\boldsymbol{\mathcal{T}} = (\mathcal{T}_1, \ldots, \mathcal{T}_n)^\top$ *is a* maximum SINR threshold feedback policy *(MTFP) if, for all* $i \in \mathcal{N}$*, there is a threshold* τ_i *such that* $\mathcal{T}_i(\gamma_i)$ *produces a feedback packet requesting beam* k *and containing* $\gamma_{k,i}$ *as the sole* SINR *information if and only if* $b_i^\star = k$ *and* $\gamma_i^\star \geq \tau_i$.

For a given set of threshold values, it is not hard to see that the GTFP (corresponding to these threshold values) always achieves a rate at least as good as the rate achieved by the MTFP (corresponding to the same threshold values) because MUs request all the beams with SINR values above their thresholds under the GTFP, which includes the best beam with the highest SINR. Since maximum SINR values are also fed back by GTFPs, they can be considered more general than MTFPs. Moreover, as shown later in Lemma 3.5, a GTFP reduces to an MTFP if threshold values of all MUs are greater than one. In this section, we will first prove that GTFPs form a rate-wise optimal subset of general decentralized feedback policies, and then obtain a similar result for MTFPs.

3.3.1 Optimality of General Threshold Feedback Policies

It is enough to focus only on the first beam since $R(\boldsymbol{\mathcal{F}})$ can be written as

$$R(\boldsymbol{\mathcal{F}}) = M\mathsf{E}_{\boldsymbol{\Gamma}}\left[\log\left(1 + \max_{i\in\mathcal{G}_1(\boldsymbol{\mathcal{F}}(\boldsymbol{\Gamma}))} \gamma_{1,i}\right)\right] \tag{3.9}$$

under our assumptions in Section 3.2. For our proofs, we will define various sets whose elements lie in various spaces including $\mathbb{R}_+^M$ and $\mathbb{R}_+^{M\times n}$. Therefore, paying attention to

the space in which the elements of a set lie will facilitate exposition in the rest of the chapter.

For a given beam symmetric general decentralized feedback policy $\boldsymbol{\mathcal{F}} = (\mathcal{F}_1, \ldots, \mathcal{F}_n)^\top$, we let

$$FB_i = \left\{ \boldsymbol{\gamma}_i \in \mathbb{R}_+^M : \mathcal{F}_i(\boldsymbol{\gamma}_i) \text{ selects beam } 1 \right\} \tag{3.10}$$

for all $i \in \mathcal{N}$. Given $\boldsymbol{\mathcal{F}}$, we construct a GTFP $\boldsymbol{\mathcal{T}}$ by choosing τ_i as

$$\mathsf{Pr}\{\gamma_{1,i} \geq \tau_i\} = \mathsf{Pr}\{\boldsymbol{\gamma}_i \in FB_i\}$$

for all $i \in \mathcal{N}$. This construction is feasible since $\gamma_{1,i}$ is assumed to have a continuous distribution function. Such a selection of $\boldsymbol{\mathcal{T}}$ leads to a fair comparison between $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{T}}$ since $\Lambda(\boldsymbol{\mathcal{F}}) = \Lambda(\boldsymbol{\mathcal{T}})$. We divide FB_i into two disjoint sets

$$\mathcal{S}_i^L = \left\{ \boldsymbol{\gamma}_i \in \mathbb{R}_+^M : \boldsymbol{\gamma}_i \in FB_i \,\&\, \gamma_{1,i} < \tau_i \right\}, \tag{3.11}$$

and

$$\mathcal{S}_i^R = \left\{ \boldsymbol{\gamma}_i \in \mathbb{R}_+^M : \boldsymbol{\gamma}_i \in FB_i \,\&\, \gamma_{1,i} \geq \tau_i \right\}. \tag{3.12}$$

Finally, we let

$$\bar{\mathcal{S}}_i^R = \left\{ \boldsymbol{\gamma}_i \in \mathbb{R}_+^M : \boldsymbol{\gamma}_i \notin FB_i \,\&\, \gamma_{1,i} \geq \tau_i \right\}. \tag{3.13}$$

We will use these sets to show $R(\boldsymbol{\mathcal{T}}) \geq R(\boldsymbol{\mathcal{F}})$.

The proof is simple for a single MU single beam communication scenario. For a particular realization of the SINR value γ_1, the same instantaneous rate is achieved by both feedback policies if they result in the same feedback decision. On the other hand, the achieved instantaneous rate will be different if only one of the policies results in a positive feedback decision. This happens either when $\gamma_1 \in \mathcal{S}_1^L$, in which case only $\boldsymbol{\mathcal{F}}$ leads to a positive feedback decision, or when $\gamma_1 \in \bar{\mathcal{S}}_1^R$, in which case only $\boldsymbol{\mathcal{T}}$ leads to a pos-

itive feedback decision. The worst case SINR on the event $\gamma_1 \in \bar{\mathcal{S}}_1^R$ is greater than the threshold value τ_1, and the best case SINR achieved by the MU on the event $\gamma_1 \in \mathcal{S}_1^L$ is less than τ_1. Therefore, the rates achieved by $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{T}}$ can be upper and lower bounded, respectively, to show that $R(\boldsymbol{\mathcal{T}}) \geq R(\boldsymbol{\mathcal{F}})$. These ideas are formally stated through the following lemma.

Lemma 3.1. *For any general decentralized feedback policy $\boldsymbol{\mathcal{F}}$, there exists a GTFP $\boldsymbol{\mathcal{T}}$ such that $\Lambda(\boldsymbol{\mathcal{F}}) = \Lambda(\boldsymbol{\mathcal{T}})$ and $R(\boldsymbol{\mathcal{T}}) \geq R(\boldsymbol{\mathcal{F}})$, where $n = 1$ and $M = 1$.*

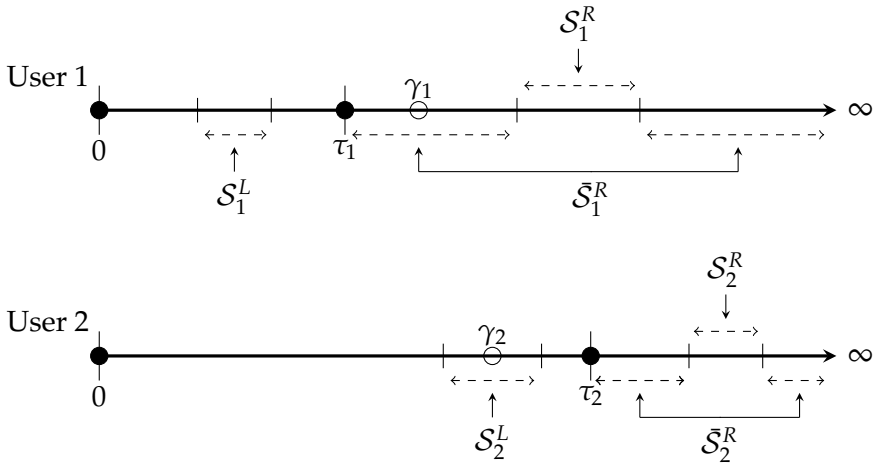
Proof. Refer Appendix 3.7.1 for the proof. □

The proof for the multiuser scenario hinges on the same principles above but it is not straightforward due to coupling effects of individual feedback rules on the aggregate rate expression. Part of the complexity to deal with these effects arises from the heterogeneous nature of the feedback rules. For example, consider a two-user single beam communication scenario. Let $\boldsymbol{\mathcal{F}} = (\mathcal{F}_1, \mathcal{F}_2)$ be a general decentralized feedback policy, and $\boldsymbol{\mathcal{T}} = (\mathcal{T}_1, \mathcal{T}_2)$ be the corresponding general threshold feedback policy as constructed above. Consider the event $\mathcal{A}$ in which $\gamma_1 \in \bar{\mathcal{S}}_1^R$ and $\gamma_2 \in \mathcal{S}_2^L$. On this event, $\boldsymbol{\mathcal{F}}$ schedules MU 2, whereas $\boldsymbol{\mathcal{T}}$ schedules MU 1. If $\tau_2 > \tau_1$, we can envisage cases in which both $\gamma_2 > \gamma_1$ and $\gamma_1 > \gamma_2$ can happen with positive probability on $\mathcal{A}$. For example, we can represent the sets of interest defined earlier on the real line in this case (*i.e.*, $M = 1$), and Figs. 3.1(a) and 3.1(b) show example realizations of γ_1 and γ_2 for which $r(\boldsymbol{\mathcal{T}}, \mathbf{\Gamma}) < r(\boldsymbol{\mathcal{F}}, \mathbf{\Gamma})$ and $r(\boldsymbol{\mathcal{T}}, \mathbf{\Gamma}) > r(\boldsymbol{\mathcal{F}}, \mathbf{\Gamma})$, respectively. Therefore, average sum-rates cannot be bound easily to determine which feedback policy achieves higher expected rate on $\mathcal{A}$. The same arguments continue to hold for other events, and the problem complexity is further magnified with increasing numbers of MUs. To overcome these issues, we will prove a more general result indicating that the best strategy for a MU is to always use a threshold feedback policy whatever the feedback policies of other MUs are.

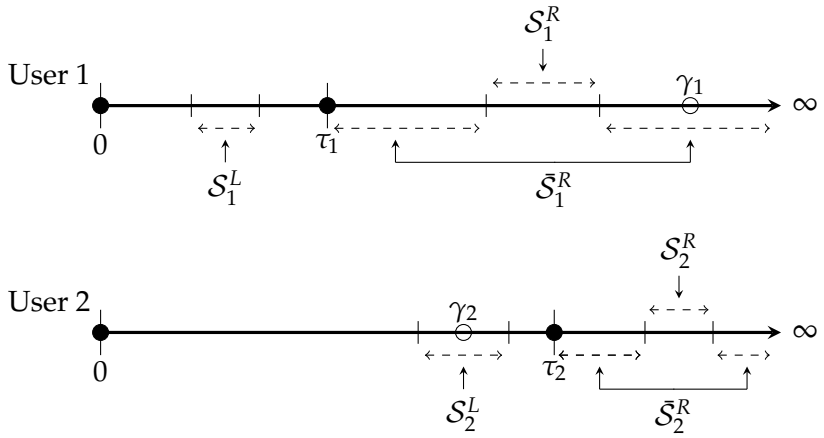
To this end, we let

$$\mathcal{G}_1^{-1}(\boldsymbol{\mathcal{F}}(\mathbf{\Gamma})) = \{i \in \mathcal{N} : i \neq 1 \,\&\, i \in \mathcal{G}_1(\boldsymbol{\mathcal{F}}(\mathbf{\Gamma}))\}$$

for a given $\boldsymbol{\mathcal{F}} = (\mathcal{F}_1, \ldots, \mathcal{F}_n)^\top$. That is, $\mathcal{G}_1^{-1}(\boldsymbol{\mathcal{F}}(\mathbf{\Gamma}))$ is the random set of MUs containing



(a) $\gamma_1 \in \bar{\mathcal{S}}_1^R$, $\gamma_2 \in \mathcal{S}_2^L$ and $r(\boldsymbol{\mathcal{T}}, \boldsymbol{\Gamma}) < r(\boldsymbol{\mathcal{F}}, \boldsymbol{\Gamma})$.



(b) $\gamma_1 \in \bar{\mathcal{S}}_1^R$, $\gamma_2 \in \mathcal{S}_2^L$ and $r(\boldsymbol{\mathcal{T}}, \boldsymbol{\Gamma}) > r(\boldsymbol{\mathcal{F}}, \boldsymbol{\Gamma})$.

Figure 3.1: A two-user example indicating problem complexity due to heterogeneity and the coupling effects between individual feedback policies.

all MUs requesting beam 1 under $\boldsymbol{\mathcal{F}}$, except for the first MU. The superscript -1 is used to indicate that all MUs but MU 1 requesting beam 1 are included in $\mathcal{G}_1^{-1}(\boldsymbol{\mathcal{F}}(\boldsymbol{\Gamma}))$. The maximum beam 1 SINR value achieved by a MU in this random set is denoted by $\bar{\gamma}_1^\star(\boldsymbol{\mathcal{F}})$, *i.e.*,

$$\bar{\gamma}_1^\star(\boldsymbol{\mathcal{F}}) = \max_{i \in \mathcal{G}_1^{-1}(\boldsymbol{\mathcal{F}}(\boldsymbol{\Gamma}))} \gamma_{1,i}.$$

Consider now the decentralized feedback policy $\boldsymbol{\mathcal{F}}^1 = (\mathcal{T}_1, \mathcal{F}_2, \cdots, \mathcal{F}_n)^\top$. That is, we

only allow MU 1 to switch to the threshold feedback rule $\mathcal{T}_1$ with the threshold value τ_1 determined as above. Then, for almost all realizations of $\mathbf{\Gamma}$, we have

$$\bar{\gamma}_1^\star(\boldsymbol{\mathcal{F}}) = \bar{\gamma}_1^\star\left(\boldsymbol{\mathcal{F}}^1\right) = \bar{\gamma}_1^\star.$$

Therefore, the difference between $R(\boldsymbol{\mathcal{F}})$ and $R\left(\boldsymbol{\mathcal{F}}^1\right)$ depends only on the rate achieved by MU 1 under these two feedback policies.

We are interested in proving $R(\boldsymbol{\mathcal{F}}) \leq R(\boldsymbol{\mathcal{T}})$. A brief sketch of the proof is as follows. We first prove that $R(\boldsymbol{\mathcal{F}}) \leq R\left(\boldsymbol{\mathcal{F}}^1\right)$. To this end, we let $\mathbf{\Gamma}_{-1}$ be the SINR matrix containing SINR values of all MUs except those of the first MU. We also let

$$R(\boldsymbol{\mathcal{F}}|\mathbf{\Gamma}_{-1}) = \mathsf{E}_{\mathbf{\Gamma}}\left[r(\boldsymbol{\mathcal{F}},\mathbf{\Gamma})\,|\mathbf{\Gamma}_{-1}\right]$$

be the conditional average sum-rate achieved by $\boldsymbol{\mathcal{F}}$ for a given $\mathbf{\Gamma}_{-1}$. Then, it is enough to show that $R\left(\boldsymbol{\mathcal{F}}^1|\mathbf{\Gamma}_{-1}\right) \geq R(\boldsymbol{\mathcal{F}}|\mathbf{\Gamma}_{-1})$ for almost all $\mathbf{\Gamma}_{-1}$. This result implies that the sum-rate increases if MU 1 switches to a threshold feedback rule regardless of feedback rules of other MUs. Repeating the same steps for other MUs $i \in \{2,3,\cdots,n\}$ one-by-one, we end up with the threshold feedback policy $\boldsymbol{\mathcal{T}}$ after n steps, and conclude that $R(\boldsymbol{\mathcal{T}}) \geq R(\boldsymbol{\mathcal{F}})$.

Before giving the details of the proof sketched above, we will first perform a preliminary analysis. For the rest of this part of the chapter, $\boldsymbol{\mathcal{F}}^1$ will represent the decentralized feedback policy derived from a given decentralized feedback policy $\boldsymbol{\mathcal{F}}$ as above. When we switch from $\boldsymbol{\mathcal{F}}$ to $\boldsymbol{\mathcal{F}}^1$, we can identify three main types of events: *neutral*, *loss* and *gain* events. On the neutral event, we will continue to achieve the same downlink throughput under both feedback policies. On the loss event, we will lose some data rate upon switching to $\boldsymbol{\mathcal{F}}^1$ from $\boldsymbol{\mathcal{F}}$. Finally, on the gain event, we will gain some data rate upon switching to $\boldsymbol{\mathcal{F}}^1$ from $\boldsymbol{\mathcal{F}}$. The difference $R\left(\boldsymbol{\mathcal{F}}^1\right) - R(\boldsymbol{\mathcal{F}})$ depends on the average sum-rates lost and gained on the loss and gain events. To show that

$$R\left(\boldsymbol{\mathcal{F}}^1\right) - R(\boldsymbol{\mathcal{F}}) \geq 0,$$

we need to characterize these loss and gain events precisely. We first formally define these events, and then provide their further characterizations suitable for our analysis in Lemmas 3.2 and 3.3.

Definition 3.5. *The loss, gain and neutral events upon switching to $\boldsymbol{\mathcal{F}}^1$ from $\boldsymbol{\mathcal{F}}$ on beam 1 are defined as*

$$A_L = \left\{ \mathbf{\Gamma} \in \mathbb{R}_+^{M \times n} \ : \ r^1 \left(\boldsymbol{\mathcal{F}}^1, \mathbf{\Gamma} \right) < r^1 \left(\boldsymbol{\mathcal{F}}, \mathbf{\Gamma} \right) \right\}, \tag{3.14}$$

$$A_G = \left\{ \mathbf{\Gamma} \in \mathbb{R}_+^{M \times n} \ : \ r^1 \left(\boldsymbol{\mathcal{F}}^1, \mathbf{\Gamma} \right) > r^1 \left(\boldsymbol{\mathcal{F}}, \mathbf{\Gamma} \right) \right\} \tag{3.15}$$

and

$$A_N = \left\{ \mathbf{\Gamma} \in \mathbb{R}_+^{M \times n} \ : \ r^1 (\boldsymbol{\mathcal{F}}^1, \mathbf{\Gamma}) = r^1 (\boldsymbol{\mathcal{F}}, \mathbf{\Gamma}) \right\}, \tag{3.16}$$

respectively.

The neutral event is not so much of an interest since both policies achieve the same rate on this event. However, loss and gain events require further evaluation, and the next two lemmas provide other characterizations for these events. These characterizations will be important when we compare $R\left(\boldsymbol{\mathcal{F}}^1\right)$ against $R\left(\boldsymbol{\mathcal{F}}\right)$.

Lemma 3.2. *A_L is equal to*

$$A_L = \left\{ \mathbf{\Gamma} \in \mathbb{R}_+^{M \times n} \ : \ \boldsymbol{\gamma}_1 \in \mathcal{S}_1^L \quad \& \quad \bar{\gamma}_1^\star < \gamma_{1,1} \right\}.$$

Proof. Refer Appendix 3.7.2 for the proof. □

A similar characterization for the gain event on beam 1 is given in the next lemma.

Lemma 3.3. *A_G is equal to*

$$A_G = \left\{ \mathbf{\Gamma} \in \mathbb{R}_+^{M \times n} \ : \ \boldsymbol{\gamma}_1 \in \bar{\mathcal{S}}_1^R \quad \& \quad \bar{\gamma}_1^\star < \gamma_{1,1} \right\}.$$

Proof. Refer Appendix 3.7.2 for the proof. □

These auxiliary results will aid to prove sum-rate optimality of $\boldsymbol{\mathcal{F}}^1$ over $\boldsymbol{\mathcal{F}}$ in Theorem 3.1. Before providing the details of the proof of this theorem, we will again give a sketch of the proof. A_L, A_G and A_N are three disjoint events with total probability mass of one. Therefore, for a feedback policy $\boldsymbol{\mathcal{F}}$, we can write

$$R^1\left(\boldsymbol{\mathcal{F}}|\boldsymbol{\Gamma}_{-1}\right) = R^1\left(\boldsymbol{\mathcal{F}}, A_L|\boldsymbol{\Gamma}_{-1}\right) + R^1\left(\boldsymbol{\mathcal{F}}, A_G|\boldsymbol{\Gamma}_{-1}\right) + R^1\left(\boldsymbol{\mathcal{F}}, A_N|\boldsymbol{\Gamma}_{-1}\right).$$

We can write a similar expression for $R^1\left(\boldsymbol{\mathcal{F}}^1|\boldsymbol{\Gamma}_{-1}\right)$. Comparison of these two expressions term-by-term reveals that

$$R^1\left(\boldsymbol{\mathcal{F}}^1|\boldsymbol{\Gamma}_{-1}\right) \geq R^1\left(\boldsymbol{\mathcal{F}}|\boldsymbol{\Gamma}_{-1}\right).$$

Since this inequality holds for almost all $\boldsymbol{\Gamma}_{-1}$, we also have

$$R^1\left(\boldsymbol{\mathcal{F}}^1\right) \geq R^1\left(\boldsymbol{\mathcal{F}}\right).$$

Since beams are statistically identical, the total rate is M times the rate achieved on beam 1. Therefore, we finally have

$$R\left(\boldsymbol{\mathcal{F}}^1\right) \geq R\left(\boldsymbol{\mathcal{F}}\right). \tag{3.17}$$

We make this idea formal in the proof of the next theorem.

Theorem 3.1. *Let*

$$\boldsymbol{\mathcal{F}} = \left(\mathcal{F}_1, \ldots, \mathcal{F}_n\right)^\top$$

and

$$\boldsymbol{\mathcal{F}}^1 = \left(\mathcal{T}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n\right)^\top$$

be defined as above. Then, $\Lambda\left(\boldsymbol{\mathcal{F}}\right) = \Lambda\left(\boldsymbol{\mathcal{F}}^1\right)$, *and* $R\left(\boldsymbol{\mathcal{F}}^1\right) \geq R\left(\boldsymbol{\mathcal{F}}\right)$ *for any* $M \geq 1$.

Proof. Refer Appendix 3.7.3 for the proof. □

This theorem shows that if a MU starts using a threshold feedback rule, the sum-rate improves regardless of the feedback rules of all other MUs. This leads to the following

key finding.

Theorem 3.2. *For any beam symmetric general decentralized feedback policy $\boldsymbol{\mathcal{F}}$, there exists a GTFP $\boldsymbol{\mathcal{T}}$ such that $\Lambda(\boldsymbol{\mathcal{F}}) = \Lambda(\boldsymbol{\mathcal{T}})$ and $R(\boldsymbol{\mathcal{T}}) \geq R(\boldsymbol{\mathcal{F}})$.*

Proof. For a given $\boldsymbol{\mathcal{F}} = (\mathcal{F}_1, \ldots, \mathcal{F}_n)^\top$, let $\boldsymbol{\mathcal{T}} = (\mathcal{T}_1, \ldots, \mathcal{T}_n)^\top$ be the GTFP constructed as above. Let

$$\boldsymbol{\mathcal{F}}^k = (\mathcal{T}_1, \ldots, \mathcal{T}_k, \mathcal{F}_{k+1}, \ldots, \mathcal{F}_n)^\top$$

for $1 \leq k \leq n-1$. When $k = n$, we have $\boldsymbol{\mathcal{F}}^n = \boldsymbol{\mathcal{T}}$. By Theorem 3.1, we have

$$R(\boldsymbol{\mathcal{F}}) \leq R\left(\boldsymbol{\mathcal{F}}^1\right) \leq \cdots \leq R(\boldsymbol{\mathcal{F}}^n) = R(\boldsymbol{\mathcal{T}}).$$

Since

$$\Lambda(\boldsymbol{\mathcal{F}}) = \Lambda\left(\boldsymbol{\mathcal{F}}^1\right) = \cdots = \Lambda(\boldsymbol{\mathcal{F}}^n) = \Lambda(\boldsymbol{\mathcal{T}}),$$

the proof is complete. □

3.3.2 Optimality of Maximum SINR Threshold Feedback Policies

In this part, we briefly explain why similar results also hold for MTFPs. The proof techniques are the same except for some subtle differences. To start with, under a maximum SINR decentralized feedback policy, each MU requests only the beam achieving the maximum SINR if the feedback conditions are met, *i.e.*, see Definitions 3.1 and 3.4. Hence, the thresholds are set such that

$$\mathsf{Pr}\{b_i^\star = 1 \text{ and } \gamma_i^\star \geq \tau_i\} = \mathsf{Pr}\{\boldsymbol{\gamma_i} \in FB_i\}.$$

The definition of FB_i is refined in which MU i requests beam 1 if and only if $b_i^\star = 1$ and $\gamma_i^\star$ satisfies feedback conditions. The definitions of other sets and events of interest require

only some subtle modifications, too. For example, A_L can now be defined as

$$A_L = \left\{ \mathbf{\Gamma} \in \mathbb{R}_+^{M \times n} \, : \, \gamma_1 \in \mathcal{S}_1^L \quad \& \quad \bar{\gamma}_1^\star < \gamma_1^\star \right\},$$

where

$$\mathcal{S}_1^L = \left\{ \gamma_1 \in \mathbb{R}_+^M : \gamma_1 \in FB_1 \, \& \, \gamma_1^\star < \tau_1 \right\}.$$

The next two theorems provide results analogous to the ones stated in Theorems 3.1 and 3.2.

Theorem 3.3. *For a given beam symmetric decentralized maximum* SINR *policy* $\boldsymbol{\mathcal{F}} = (\mathcal{F}_1, \ldots, \mathcal{F}_n)^\top$, *let* $\boldsymbol{\mathcal{F}}^1 = (\mathcal{T}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n)^\top$ *be the maximum* SINR *threshold feedback policy derived from* $\boldsymbol{\mathcal{F}}$ *by allowing MU* 1 *to switch from* $\mathcal{F}_1$ *to* $\mathcal{T}_1$, *where* $\mathcal{T}_1$ *is a beam symmetric maximum* SINR *threshold rule whose threshold is set as above. Then,* $\Lambda(\boldsymbol{\mathcal{F}}) = \Lambda\left(\boldsymbol{\mathcal{F}}^1\right)$, *and* $R\left(\boldsymbol{\mathcal{F}}^1\right) \geq R(\boldsymbol{\mathcal{F}})$ *for any* $M \geq 1$.

Theorem 3.4. *For any beam symmetric decentralized maximum* SINR *feedback policy* $\boldsymbol{\mathcal{F}}$, *there exists an MTFP* $\boldsymbol{\mathcal{T}}$ *such that* $\Lambda(\boldsymbol{\mathcal{F}}) = \Lambda(\boldsymbol{\mathcal{T}})$ *and* $R(\boldsymbol{\mathcal{T}}) \geq R(\boldsymbol{\mathcal{F}})$.

Since the proofs of these theorems are similar to the proofs above, we skip them to avoid repetition. It is important to note that Theorems 3.2 and 3.4 hold for any continuous SINR distribution.

3.4 Special Case: Homogenous Threshold Feedback Policies

Due to the general nature of the analysis, we can directly extend the results in the previous sections for a case where the MUs use a homogenous feedback policy. However, in this section, we give a simpler proof using basic set theory for such a scenario. We will use the same sets FB, $\mathcal{S}^L$, $\mathcal{S}^R$ and $\bar{\mathcal{S}}^R$ given in (3.10), (3.11), (3.12), and (3.13), respectively, but we will drop the user index since it will be unnecessary due to the homogeneity. Again, if the best MU has a positive feedback decision on both the feedback policies, the system will achieve the same instantaneous downlink throughput. Therefore, any event

which creates a positive feedback decision at the best MU on one feedback policy, but not on the other, creates interest. As a start, let's define the event

$$A_S = \left(\bigcup_{i=1}^{n}\left\{\gamma_i \in \mathcal{S}^L\right\}\right) \bigcap \left(\bigcap_{i=1}^{n}\left\{\gamma_i \notin \mathcal{S}^R\right\}\right),$$

which is true when there exists at least one MU having a SINR vector from the set $\mathcal{S}^L$, and there is no MU having a SINR vector from $\mathcal{S}^R$. We define another event

$$A_\tau = \left(\bigcup_{i=1}^{n}\left\{\gamma_i \in \bar{\mathcal{S}}^R\right\}\right) \bigcap \left(\bigcap_{i=1}^{n}\left\{\gamma_i \notin \mathcal{S}^R\right\}\right),$$

which is true when there exists at least one MU having a SINR vector from the set $\bar{\mathcal{S}}^R$, and there is no MU having a SINR vector from $\mathcal{S}^R$. Note that in both the events, the possibility of a MU getting a SINR vector from $\mathcal{S}^R$ is eliminated. This is done to avoid the neutral event. Also note that,

$$\mathsf{Pr}\left\{A_S\right\} = \mathsf{Pr}\left\{A_\tau\right\} = \sum_{i=1}^{n}\binom{n}{i} p_L^i (1 - \mathsf{Pr}\left\{\gamma_{1,i} \geq \tau\right\})^{n-i} = p_{ev},$$

where $p_L = \mathsf{Pr}\left\{\gamma_i \in \mathcal{S}^L\right\}$, and we also have $p_L = \mathsf{Pr}\left\{\gamma_i \in \bar{\mathcal{S}}^R\right\}$ from the definition of the sets.

Now, let's look at the complements of these events which are stated in the following lemma.

Lemma 3.4. *The complement of A_S is given by*

$$A_S^c = \left(\bigcap_{i=1}^{n}\left\{\gamma_i \notin FB\right\}\right) \bigcup \left(\bigcup_{i=1}^{n}\left\{\gamma_i \in \mathcal{S}^R\right\}\right),$$

and the complement of A_τ is

$$A_\tau^c = \left(\bigcap_{i=1}^{n}\left\{\gamma_{1,i} < \tau\right\}\right) \bigcup \left(\bigcup_{i=1}^{n}\left\{\gamma_i \in \mathcal{S}^R\right\}\right).$$

Proof. Refer Appendix 3.7.4 for the proof. □

In the proof of the next theorem, these auxiliary results are used to prove the optimality.

Theorem 3.5. *For any beam symmetric decentralized homogenous feedback policy $\boldsymbol{\mathcal{F}}$, there exists a homogenous threshold feedback policy $\boldsymbol{\mathcal{T}}$ such that $\Lambda(\boldsymbol{\mathcal{F}}) = \Lambda(\boldsymbol{\mathcal{T}})$ and $R(\boldsymbol{\mathcal{T}}) \geq R(\boldsymbol{\mathcal{F}})$.*

Proof. Refer Appendix 3.7.5 for the proof. □

3.5 Discussion of Results

In this part, we briefly discuss the results presented above. We start with a comparison between GTFPs and MTFPs. The main advantage of GTFPs over MTFPs is the ability of the base station to allocate multiple beams to a MU. Therefore, a GTFP policy achieves higher data rates when compared to an MTFP policy with the same threshold levels. From a practical point of view, such gains in data rates are expected to be minor due to dependencies among beams at a MU, *i.e.*, high $\gamma_{m,i}$ implies low $\gamma_{k,i}$, $\forall k \neq m$. Moreover, both types of policies achieve the same performance if all threshold values are greater than 1, which is formally proved in the next lemma.

Lemma 3.5. *Let $\boldsymbol{\mathcal{T}}$ be an MTFP with thresholds $\{\tau_i\}_{i\in\mathcal{N}}$, and $\boldsymbol{\mathcal{T}}'$ be the corresponding GTFP with the same threshold levels. Let $\mathcal{N}_m$ and $\mathcal{N}'_m$ be the sets of MUs requesting beam $m \in \mathcal{M}$ according to $\boldsymbol{\mathcal{T}}$ and $\boldsymbol{\mathcal{T}}'$, respectively. If $\tau_i > 1$ for all $i \in \mathcal{N}$, then $\mathcal{N}_m = \mathcal{N}'_m$.*

Proof. Refer Appendix 3.7.6 for the proof. □

Note that the requirement on threshold values for the equality of MTFPs and GTFPs in Lemma 3.5 is only a 0 [dB] requirement, which is practically a quite low SINR value. This implies that both feedback policies will actually achieve the same sum-rate in almost all practical communication scenarios.

On the other hand, from a theoretical point of view, the resulting optimization problem over $\mathbb{R}^n_+$ lends itself more amenable to further analysis if we only focus on GTFPs. More specifically, we can search for the optimal beam symmetric feedback policies within the class of GTFPs without sacrificing from optimality thanks to Theorem 3.2, and with a

slight abuse of notation, we can equivalently write (3.8) as

$$\begin{array}{ll} \underset{\boldsymbol{\tau}\in\mathbb{R}_+^n}{\text{maximize}} & R\left(\boldsymbol{\tau}\right) \\ \text{subject to} & \sum_{i=1}^{n} \mathsf{Pr}\left\{\gamma_{1,i} \geq \tau_i\right\} \leq \lambda \end{array}. \tag{3.18}$$

Some further game theoretic insights are as follows. We will only focus on GTFPs but similar explanations also hold for MTFPs. Given the same utility function $R\left(\mathcal{F}_1, \ldots, \mathcal{F}_n\right)$ for all MUs, the selfish optimization problem faced by MU i is to choose a beam symmetric decentralized feedback rule maximizing its utility given other MUs' feedback rules without increasing the feedback level. Theorem 3.1 shows that the *dominant strategy* is to switch from $\mathcal{F}_i$ to the corresponding threshold rule $\mathcal{T}_i$. As a result, the set of GTFPs constitute the set of Nash equilibria for this feedback rule selection game, and therefore GTFPs are also stable operating points from a game theoretic point of view.

In the next chapter, we will analyze the finite dimensional optimization problem in (3.18). We will show that the sum-rate becomes a Schur-concave function of feedback probabilities $p_i = \mathsf{Pr}\left\{\gamma_{1,i} \geq \tau_i\right\}$ if the SINR distribution satisfies some mild conditions. This result establishes the optimality of homogenous general threshold feedback policies among the class of beam symmetric general decentralized feedback policies.

3.6 Conclusions

Opportunistic beamforming is an important communication strategy achieving the full CSI sum-rate capacity for vector broadcast channels to a first order by only requiring partial CSI at the base station. Nevertheless, it cannot eliminate the linear growth in the feedback load with increasing numbers of MUs in the network unless a selective feedback policy is implemented for user selection. In this chapter, we have been motivated by these considerations to analyze the resulting downlink sum-rate with user selection when orthonormal beams are opportunistically allocated to MUs for the downlink communication. In particular, we have focused on the structure of optimal selective decentralized feedback policies for opportunistic beamforming under finite feedback constraints on the average number of MUs feeding back. We have shown that threshold feedback policies

in which MUs compare their beam SINRs with a threshold for their feedback decisions are always optimal to maximize the downlink sum-rate. This class of policies was studied in many previous works such as [21,29,34,70,73,79] without any formal justification for why they are the right choice for user selection. Our thresholding optimality result provides the formal justification, which holds for all fading channel models with continuous distribution functions. Since each threshold feedback policy can be associated with a threshold vector in $\mathbb{R}_+^n$, these results also reduce the search for rate-wise optimal feedback policies from function spaces to finite dimensional Euclidean spaces, which will be analyzed in the next chapter.

3.7 Appendix

3.7.1 Proof of Lemma 3.1

Using the definition of rate, the average throughput achieved using $\boldsymbol{\mathcal{F}}$ is given by

$$R\left(\boldsymbol{\mathcal{F}}\right) = \mathsf{E}\left[\log(1+\gamma_1).1_{\left\{\gamma_1 \in \mathcal{S}_1^L\right\}}\right] + \mathsf{E}\left[\log(1+\gamma_1).1_{\left\{\gamma_1 \in \mathcal{S}_1^R\right\}}\right],$$

and since $\sup \mathcal{S}_1^L$ is less than the threshold,

$$R\left(\boldsymbol{\mathcal{F}}\right) \leq \mathsf{Pr}\left\{\gamma_1 \in \mathcal{S}_1^L\right\} \log(1+\tau_1) + \mathsf{E}\left[\log(1+\gamma_1).1_{\left\{\gamma_1 \in \mathcal{S}_1^R\right\}}\right].$$

Similarly, the average throughput achieved using $\boldsymbol{\mathcal{T}}$ can be written as

$$R\left(\boldsymbol{\mathcal{T}}\right) = \mathsf{E}\left[\log(1+\gamma_1).1_{\left\{\gamma_1 \in \bar{\mathcal{S}}_1^R\right\}}\right] + \mathsf{E}\left[\log(1+\gamma_1).1_{\left\{\gamma_1 \in \mathcal{S}_1^R\right\}}\right],$$

which can be lower bounded as,

$$R\left(\boldsymbol{\mathcal{T}}\right) \geq \mathsf{Pr}\left\{\gamma_1 \in \bar{\mathcal{S}}_1^R\right\} \log(1+\tau_1) + \mathsf{E}\left[\log(1+\gamma_1).1_{\{\gamma_1 \in \mathcal{S}^R\}}\right] = R\left(\boldsymbol{\mathcal{F}}\right)$$

since

$$\mathsf{Pr}\left\{\gamma_1 \in \bar{\mathcal{S}}_1^R\right\} = \mathsf{Pr}\left\{\gamma_1 \in \mathcal{S}_1^L\right\},$$

which completes the proof.

3.7.2 Loss Event and Gain Event

Proof of Lemma 3.2

Set

$$\bar{A}_L = \left\{\mathbf{\Gamma} \in \mathbb{R}_+^{M\times n} \,:\, \boldsymbol{\gamma}_1 \in \mathcal{S}_1^L \,\&\, \bar{\gamma}_1^\star < \gamma_{1,1}\right\}.$$

We will show $\bar{A}_L = A_L$. For all $\mathbf{\Gamma}$ with $\boldsymbol{\gamma}_1 \in \mathcal{S}_1^L$, MU 1 requests beam 1 under $\boldsymbol{\mathcal{F}}$, but not under $\boldsymbol{\mathcal{F}}^1$. Therefore, if $\boldsymbol{\gamma}_1 \in \mathcal{S}_1^L$ and $\bar{\gamma}_1^\star < \gamma_{1,1}$, the system using $\boldsymbol{\mathcal{F}}$ schedules MU 1 for communication along beam 1, and the system using $\boldsymbol{\mathcal{F}}^1$ schedules another MU having $\bar{\gamma}_1^\star < \gamma_{1,1}$ for communication along beam 1. This means $r^1\left(\boldsymbol{\mathcal{F}}^1, \mathbf{\Gamma}\right) < r^1\left(\boldsymbol{\mathcal{F}}, \mathbf{\Gamma}\right)$ for all $\mathbf{\Gamma} \in \bar{A}_L$, implying $\bar{A}_L \subseteq A_L$.

Showing $A_L \subseteq \bar{A}_L$ will complete the proof. For all $\mathbf{\Gamma}$ with $\bar{\gamma}_1^\star \geq \gamma_{1,1}$, both feedback policies will achieve the same throughput by scheduling the MU having $\bar{\gamma}_1^\star$. Therefore, we must have $\bar{\gamma}_1^\star < \gamma_{1,1}$ on the loss event. Now, if $\boldsymbol{\gamma}_1 \notin FB_1$, MU 1 will not feed back under $\boldsymbol{\mathcal{F}}$, which implies no potential loss on beam 1. Therefore, for all $\mathbf{\Gamma} \in A_L$, we must have $\boldsymbol{\gamma}_1 \in FB_1$ and $\bar{\gamma}_1^\star < \gamma_{1,1}$. If $\bar{\gamma}_1^\star < \gamma_{1,1}$ and $\boldsymbol{\gamma}_1 \in \mathcal{S}_1^R$, MU 1 requests beam 1 under both feedback policies, resulting in a neutral event. This implies that $\boldsymbol{\gamma}_1 \in \mathcal{S}_1^L$ and $\bar{\gamma}_1^\star < \gamma_{1,1}$ for all $\mathbf{\Gamma} \in A_L$. Therefore, we also have $A_L \subseteq \bar{A}_L$, which concludes the proof.

Proof of Lemma 3.3

The proof is similar to the one given for Lemma 3.2. Set

$$\bar{A}_G = \left\{\mathbf{\Gamma} \in \mathbb{R}_+^{M\times n} \,:\, \boldsymbol{\gamma}_1 \in \bar{\mathcal{S}}_1^R \,\&\, \bar{\gamma}_1^\star < \gamma_{1,1}\right\}.$$

We first show that $\bar{A}_G \subseteq A_G$. For all $\mathbf{\Gamma}$ with $\boldsymbol{\gamma}_1 \in \bar{\mathcal{S}}_1^R$ and $\bar{\gamma}_1^\star < \gamma_{1,1}$, a system using $\boldsymbol{\mathcal{F}}^1$ schedules MU 1 for communication on beam 1, but a system using $\boldsymbol{\mathcal{F}}$ schedules the MU with $\bar{\gamma}_1^\star < \gamma_{1,1}$. Therefore, $r_1\left(\boldsymbol{\mathcal{F}}^1, \mathbf{\Gamma}\right) > r_1\left(\boldsymbol{\mathcal{F}}, \mathbf{\Gamma}\right)$ if $\boldsymbol{\gamma}_1 \in \bar{\mathcal{S}}_1^R$ and $\bar{\gamma}_1^\star < \gamma_{1,1}$, implying $\bar{A}_G \subseteq A_G$.

Next, observe that the neutral event occurs for all $\mathbf{\Gamma}$ with $\bar{\gamma}_1^\star \geq \gamma_{1,1}$. Therefore, we must have $\bar{\gamma}_1^\star < \gamma_{1,1}$ on the gain event. If $\gamma_{1,1} < \tau_1$, MU 1 will not feed back under $\boldsymbol{\mathcal{F}}^1$, and therefore no rate gain is achieved by switching to $\boldsymbol{\mathcal{F}}^1$. Therefore, we must have $\gamma_{1,1} \geq \tau_1$ on the gain event. If $\boldsymbol{\gamma}_1 \in \mathcal{S}_1^R$, MU 1 still feeds back under both feedback policies, which again leads to a neutral event. Therefore, for all $\mathbf{\Gamma} \in A_G$, we must have $\boldsymbol{\gamma}_1 \in \bar{\mathcal{S}}_1^R$ and $\bar{\gamma}_1^\star < \gamma_{1,1}$, which shows that $A_G \subseteq \bar{A}_G$ and completes the proof.

3.7.3 Proof of Theorem 3.1

It is enough to prove $R^1\left(\boldsymbol{\mathcal{F}}^1 | \mathbf{\Gamma}_{-1}\right) \geq R^1\left(\boldsymbol{\mathcal{F}} | \mathbf{\Gamma}_{-1}\right)$ for almost all $\mathbf{\Gamma}_{-1}$. By definition, we have

$$R^1\left(\boldsymbol{\mathcal{F}}, A_N | \mathbf{\Gamma}_{-1}\right) = R^1\left(\boldsymbol{\mathcal{F}}^1, A_N | \mathbf{\Gamma}_{-1}\right),$$

and therefore we are only interested in the average sum-rates on loss and gain events.

The following identity follows from the definition of conditional expectation.

$$R^1\left(\boldsymbol{\mathcal{F}}, A_L | \mathbf{\Gamma}_{-1}\right) = \mathsf{Pr}\left(A_L | \mathbf{\Gamma}_{-1}\right) \mathsf{E}_{\mathbf{\Gamma}}\left[r^1\left(\boldsymbol{\mathcal{F}}, \mathbf{\Gamma}\right) | A_L, \mathbf{\Gamma}_{-1}\right].$$

Lemma 3.2 implies that whenever A_L is correct, MU 1 requests beam 1, and achieves the best SINR on beam 1 among all the MUs requesting beam 1. Since $\boldsymbol{\gamma}_1 \in \mathcal{S}_1^L$ on A_L, $\gamma_{1,1}$ is less than τ_1. Therefore,

$$R^1\left(\boldsymbol{\mathcal{F}}, A_L | \mathbf{\Gamma}_{-1}\right) \quad \leq \quad \mathsf{Pr}\left(A_L | \mathbf{\Gamma}_{-1}\right) \log\left(1 + \tau_1\right). \tag{3.19}$$

Similarly, we can write

$$R^1\left(\boldsymbol{\mathcal{F}}, A_G | \mathbf{\Gamma}_{-1}\right) = \mathsf{Pr}\left(A_G | \mathbf{\Gamma}_{-1}\right) \mathsf{E}_{\mathbf{\Gamma}}\left[r^1\left(\boldsymbol{\mathcal{F}}, \mathbf{\Gamma}\right) | A_G, \mathbf{\Gamma}_{-1}\right].$$

Lemma 3.3 implies that MU 1 achieves the best SINR on beam 1 among all the MUs requesting beam 1 but $\gamma_1 \in \bar{\mathcal{S}}_1^R$ on A_G. Therefore, $\gamma_1 \notin FB_1$, and MU 1 will not request beam 1 under $\boldsymbol{\mathcal{F}}$. Hence, $\boldsymbol{\mathcal{F}}$ schedules beam 1 to the MU with SINR value $\bar{\gamma}_1^\star$, which leads to [4]

$$R^1\left(\boldsymbol{\mathcal{F}}, A_G|\boldsymbol{\Gamma}_{-1}\right) \;=\; \mathsf{Pr}\left(A_G|\boldsymbol{\Gamma}_{-1}\right)\log\left(1+\bar{\gamma}_1^\star\right). \tag{3.20}$$

Similar to the above arguments, MU 1 will not request beam 1 under $\boldsymbol{\mathcal{F}}^1$ on the event A_L since $\gamma_1 \in \mathcal{S}_1^L$. This means

$$R^1\left(\boldsymbol{\mathcal{F}}^1, A_L|\boldsymbol{\Gamma}_{-1}\right) = \mathsf{Pr}\left(A_L|\boldsymbol{\Gamma}_{-1}\right)\log\left(1+\bar{\gamma}_1^\star\right). \tag{3.21}$$

Finally, MU 1 requests beam 1 under $\boldsymbol{\mathcal{F}}^1$ on A_G, leading to

$$R^1\left(\boldsymbol{\mathcal{F}}^1, A_G|\boldsymbol{\Gamma}_{-1}\right) \geq \mathsf{Pr}\left(A_G|\boldsymbol{\Gamma}_{-1}\right)\log\left(1+\max\left(\tau_1, \bar{\gamma}_1^\star\right)\right). \tag{3.22}$$

By using (3.19), (3.20), (3.21) and (3.22), we have

$$\begin{aligned} R^1\left(\boldsymbol{\mathcal{F}}^1|\boldsymbol{\Gamma}_{-1}\right) - R^1\left(\boldsymbol{\mathcal{F}}|\boldsymbol{\Gamma}_{-1}\right) \geq\; & \mathsf{Pr}\left(A_G|\boldsymbol{\Gamma}_{-1}\right)\left(\log\left(1+\max\left(\tau_1, \bar{\gamma}_1^\star\right)\right) - \log\left(1+\bar{\gamma}_1^\star\right)\right) \\ & + \mathsf{Pr}\left(A_L|\boldsymbol{\Gamma}_{-1}\right)\left(\log\left(1+\bar{\gamma}_1^\star\right) - \log\left(1+\tau_1\right)\right). \end{aligned}$$

To conclude the proof, we need to analyze two different cases separately. If $\bar{\gamma}_1^\star \geq \tau_1$, then it directly follows that

$$R^1\left(\boldsymbol{\mathcal{F}}^1|\boldsymbol{\Gamma}_{-1}\right) - R^1\left(\boldsymbol{\mathcal{F}}|\boldsymbol{\Gamma}_{-1}\right) \geq 0.$$

If $\bar{\gamma}_1^\star < \tau_1$, then we have

$$\begin{aligned} & R^1\left(\boldsymbol{\mathcal{F}}^1|\boldsymbol{\Gamma}_{-1}, \bar{\gamma}_1^\star < \tau_1\right) - R^1\left(\boldsymbol{\mathcal{F}}|\boldsymbol{\Gamma}_{-1}, \bar{\gamma}_1^\star < \tau_1\right) \\ & \qquad \geq \left(\mathsf{Pr}\left(A_G|\boldsymbol{\Gamma}_{-1}, \bar{\gamma}_1^\star < \tau_1\right) - \mathsf{Pr}\left(A_L|\boldsymbol{\Gamma}_{-1}, \bar{\gamma}_1^\star < \tau_1\right)\right)\left(\log\left(1+\tau_1\right) - \log\left(1+\bar{\gamma}_1^\star\right)\right). \end{aligned}$$

[4] Note that $\bar{\gamma}_1^\star$ is a (measurable) function of $\boldsymbol{\Gamma}_{-1}$, and therefore (3.20) conforms with the measure theoretic definition of the conditional expectation.

Observe that

$$\mathsf{Pr}\left(A_G|\mathbf{\Gamma}_{-1},\bar{\gamma}_1^\star<\tau_1\right)=\mathsf{Pr}\left\{\gamma_1\in\bar{\mathcal{S}}_1^R\right\}$$

and

$$\mathsf{Pr}\left\{A_L|\mathbf{\Gamma}_{-1},\bar{\gamma}_1^\star<\tau_1\right\}\leq\mathsf{Pr}\left\{\gamma_1\in\mathcal{S}_1^L\right\}.$$

Since $\mathsf{Pr}\left\{\gamma_1\in\bar{\mathcal{S}}_1^R\right\}=\mathsf{Pr}\left\{\gamma_1\in\mathcal{S}_1^L\right\}$, we have

$$R^1\left(\boldsymbol{\mathcal{F}}^1|\mathbf{\Gamma}_{-1},\bar{\gamma}_1^\star<\tau_1\right)-R^1\left(\boldsymbol{\mathcal{F}}|\mathbf{\Gamma}_{-1},\bar{\gamma}_1^\star<\tau_1\right)\geq 0.$$

After removing conditioning, this proves that $R^1\left(\boldsymbol{\mathcal{F}}^1|\mathbf{\Gamma}_{-1}\right)\geq R^1\left(\boldsymbol{\mathcal{F}}|\mathbf{\Gamma}_{-1}\right)$ for almost all $\mathbf{\Gamma}_{-1}$, and therefore

$$R^1\left(\boldsymbol{\mathcal{F}}^1\right)\geq R^1\left(\boldsymbol{\mathcal{F}}\right).$$

3.7.4 Proof of Lemma 3.4

Using set theory, we have

$$A_S^c=\left(\bigcap_{i=1}^n\left\{\gamma_i\notin\mathcal{S}^L\right\}\right)\bigcup\left(\bigcup_{i=1}^n\left\{\gamma_i\in\mathcal{S}^R\right\}\right).$$

Since the complement of $\bigcap_{i=1}^n\{\gamma_i\notin FB\}$ is $\bigcup_{i=1}^n\{\gamma_i\in FB\}$, we have

$$A_S^c=\left(\bigcap_{i=1}^n\left\{\gamma_i\notin\mathcal{S}^L\right\}\right)\bigcap\left(\left\{\bigcap_{i=1}^n\{\gamma_i\notin FB\}\right\}\bigcup\left\{\bigcup_{i=1}^n\{\gamma_i\in FB\}\right\}\right)\bigcup\left(\bigcup_{i=1}^n\left\{\gamma_i\in\mathcal{S}^R\right\}\right).$$

This can be rewritten as

$$A_S^c=\left(\bigcap_{i=1}^n\left\{\gamma_i\notin\mathcal{S}^L,FB\right\}\right)\bigcup\left(\left\{\bigcup_{i=1}^n\{\gamma_i\in FB\}\right\}\bigcap\left\{\bigcap_{i=1}^n\left\{\gamma_i\notin\mathcal{S}^L\right\}\right\}\right)$$

$$\bigcup \left(\bigcup_{i=1}^{n} \left\{ \gamma_i \in \mathcal{S}^R \right\} \right).$$

Since

$$\left\{ \bigcup_{i=1}^{n} \{ \gamma_i \in FB \} \right\} \bigcap \left\{ \bigcap_{i=1}^{n} \left\{ \gamma_i \notin \mathcal{S}^L \right\} \right\} = \bigcup_{i=1}^{n} \left\{ \gamma_i \in \mathcal{S}^R \right\},$$

we have

$$A_S^c = \left(\bigcap_{i=1}^{n} \{ \gamma_i \notin FB \} \right) \bigcup \left(\bigcup_{i=1}^{n} \left\{ \gamma_i \in \mathcal{S}^R \right\} \right).$$

A_τ^c can be obtained by following the same lines of the proof of A_S^c, which completes this proof.

3.7.5 Proof of Theorem 3.5

We can write the average rate on beam one as

$$\begin{aligned} R^1(\boldsymbol{\mathcal{F}}) &= R^1(\boldsymbol{\mathcal{F}}, A_S) + R^1(\boldsymbol{\mathcal{F}}, A_S^c) \\ &= \mathsf{E}\left[\log\left(1 + \max_{1 \le i \le n} \gamma_{1,i} 1_{\{\gamma_i \in FB\}} \right) 1_{\{(\bigcup_{i=1}^n \{\gamma_i \in \mathcal{S}^L\}) \bigcap (\bigcap_{i=1}^n \{\gamma_i \notin \mathcal{S}^R\})\}} \right] \\ &+ \mathsf{E}\left[\log\left(1 + \max_{1 \le i \le n} \gamma_{1,i} 1_{\{\gamma_i \in FB\}} \right) 1_{\{(\bigcap_{i=1}^n \{\gamma_i \notin FB\}) \bigcup (\bigcup_{i=1}^n \{\gamma_i \in \mathcal{S}^R\})\}} \right]. \end{aligned}$$

Since $\bigcap_{i=1}^n \{\gamma_i \notin FB\}$ and $\bigcup_{i=1}^n \{\gamma_i \in \mathcal{S}^R\}$ are disjoint sets, we can write

$$\begin{aligned} R^1(\boldsymbol{\mathcal{F}}) = \mathsf{E}\left[\log\left(1 + \max_{1 \le i \le n} \gamma_{1,i} 1_{\{\gamma_i \in FB\}} \right) 1_{\{(\bigcup_{i=1}^n \{\gamma_i \in \mathcal{S}^L\}) \bigcap (\bigcap_{i=1}^n \{\gamma_i \notin \mathcal{S}^R\})\}} \right] \\ + \underbrace{\mathsf{E}\left[\log\left(1 + \max_{1 \le i \le n} \gamma_{1,i} 1_{\{\gamma_i \in FB\}} \right) 1_{\{\bigcap_{i=1}^n \{\gamma_i \notin FB\}\}} \right]}_{=0} \\ + \mathsf{E}\left[\log\left(1 + \max_{1 \le i \le n} \gamma_{1,i} 1_{\{\gamma_i \in FB\}} \right) . 1_{\{\bigcup_{i=1}^n \{\gamma_i \in \mathcal{S}^R\}\}} \right]. \end{aligned}$$

We will upper bound $R^1(\boldsymbol{\mathcal{F}})$ as,

$$R^1(\boldsymbol{\mathcal{F}}) \leq p_{ev}\log(1+\tau) + \mathsf{E}\left[\log\left(1+\max_{1\leq i\leq n}\gamma_{1,i}1_{\{\gamma_i\in FB\}}\right)1_{\{\bigcup_{i=1}^n\{\gamma_i\in\mathcal{S}^R\}\}}\right]$$

by evaluating the first term in the right hand side of the equation, and using the fact that there is at least one MU with a SINR vector from $\mathcal{S}^L$, and $\sup\mathcal{S}^L = \tau$. When it comes to the last term in the right hand side, it is not hard to see that if there exists a MU i with $\gamma_i 1_{\{\gamma_i\in FB\}} \in \mathcal{S}^R$, then $\max_{1\leq i\leq n}\gamma_{1,i} \in \mathcal{S}^R$. Therefore, we have

$$R^1(\boldsymbol{\mathcal{F}}) \leq p_{ev}\log(1+\tau) + \mathsf{E}\left[\log\left(1+\max_{1\leq i\leq n}\gamma_{1,i}1_{\{\gamma_i\in\mathcal{S}^R\}}\right)1_{\{\bigcup_{i=1}^n\{\gamma_i\in\mathcal{S}^R\}\}}\right].$$

Now, by using a similar technique, we will obtain a lower bound for $R^1(\boldsymbol{\mathcal{T}})$, and show that it is greater than or equal to this upper bound of $R^1(\boldsymbol{\mathcal{F}})$.

Using A_τ and A_τ^c, we have

$$R^1(\boldsymbol{\mathcal{T}}) = R^1(\boldsymbol{\mathcal{T}}, A_\tau) + R^1(\boldsymbol{\mathcal{T}}, A_\tau^c).$$

Using similar arguments to the ones done above, we have

$$R^1(\boldsymbol{\mathcal{T}}) \geq p_{ev}\log(1+\tau) + \underbrace{\mathsf{E}\left[\log\left(1+\max_{1\leq i\leq n}\gamma_{1,i}1_{\{\gamma_{1,i}\geq\tau\}}\right)1_{\{\bigcap_{i=1}^n\gamma_{1,i}<\tau\}}\right]}_{=0} + \mathsf{E}\left[\log\left(1+\max_{1\leq i\leq n}\gamma_{1,i}1_{\{\gamma_{1,i}\geq\tau\}}\right)1_{\{\bigcup_{i=1}^n\{\gamma_i\in\mathcal{S}^R\}\}}\right].$$

It is not hard to see that,

$$\mathsf{E}\left[\log\left(1+\max_{1\leq i\leq n}\gamma_{1,i}1_{\{\gamma_{1,i}\geq\tau\}}\right)1_{\{\bigcup_{i=1}^n\{\gamma_i\in\mathcal{S}^R\}\}}\right] \geq \mathsf{E}\left[\log\left(1+\max_{1\leq i\leq n}\gamma_{1,i}1_{\{\gamma_i\in\mathcal{S}^R\}}\right)1_{\{\bigcup_{i=1}^n\{\gamma_i\in\mathcal{S}^R\}\}}\right].$$

Therefore, $R(\boldsymbol{\mathcal{T}}) \geq R(\boldsymbol{\mathcal{F}})$, which completes the proof.

3.7.6 Proof of Lemma 3.5

$\mathcal{N}_m$ and $\mathcal{N}'_m$ are given as

$$\mathcal{N}_m \;=\; \left\{ i \in \mathcal{N} : b_i^* = m \;\&\; \gamma_{b_i^*,i} \geq \tau_i \right\}$$

and

$$\mathcal{N}'_m \;=\; \left\{ i \in \mathcal{N} : \gamma_{m,i} \geq \tau_i \right\}.$$

Thus, we have $\mathcal{N}_m \subseteq \mathcal{N}'_m$. To show the other direction, take any $i \in \mathcal{N}'_m$, and a beam index $r \neq m$. Then, $|\boldsymbol{h}_i^\top \boldsymbol{q}_m|^2 > |\boldsymbol{h}_i^\top \boldsymbol{q}_r|^2$ because $\tau_i > 1$. Therefore, the following holds.

$$\begin{aligned} \gamma_{m,i} &= \frac{|\boldsymbol{h}_i^\top \boldsymbol{q}_m|^2}{\frac{1}{\rho} + \sum_{k=1,k\neq m}^{M} |\boldsymbol{h}_i^\top \boldsymbol{q}_k|^2} \\ &> \frac{|\boldsymbol{h}_i^\top \boldsymbol{q}_r|^2}{\frac{1}{\rho} + \sum_{k=1,k\neq r}^{M} |\boldsymbol{h}_i^\top \boldsymbol{q}_k|^2} = \gamma_{r,i}. \end{aligned}$$

As a result, any MU $i \in \mathcal{N}'_m$ achieves its maximum SINR at beam m if $\tau_i > 1$. This implies that $b_i^* = m$ and $i \in \mathcal{N}_m$.

Chapter 4
Optimal Threshold Selection Problem

Threshold feedback policies are well known and provably rate-wise optimal selective feedback techniques for communication systems requiring partial channel state information (CSI). However, optimal selection of thresholds at mobile users to maximize information theoretic data rates subject to feedback constraints is an open problem. In this chapter, we focus on the optimal threshold selection problem, and provide a solution for this problem for finite feedback systems. Rather surprisingly, we show that using the same threshold values at all mobile users is not always a rate-wise optimal feedback strategy, even for a system with identical users experiencing statistically the same channel conditions. By utilizing the theory of majorization, we identify an underlying Schur-concave structure in the rate function and obtain sufficient conditions for a homogenous threshold feedback policy to be optimal. Our results hold for most fading channel models, and we illustrate applications of our results to well known fading channel models such as Rayleigh, Nakagami and Rician fading channels, along with various engineering and design insights. We show, for the Rayleigh fading channel model, homogenous threshold feedback policies are proven to be rate-wise optimal if multiple orthonormal data carrying beams are used to communicate with multiple mobile users simultaneously.

4.1 Introduction

CONSIDER a multiuser communication system in a fading environment. The channel changes over time, and the goal of the base station is to maximize downlink data rates by taking channel variations into account. The base station has multiple transmission antennas, and therefore can possibly communicate with a selected subset of mobile users (MUs) through multiple traffic flows simultaneously. In this setting, selecting the MUs with the best instantaneous signal-to-interference-plus-noise-ratios (SINR) for communication is a simple communication strategy that is heuristically expected to max-

imize downlink data rates. Indeed, this is the classical opportunistic communication approach. However, regardless of partial or perfect CSI feedback, all MUs contend for the uplink to communicate their CSI to the base station in such a setting. This is an impractical burden on the uplink (for large numbers of MUs), and a significant waste of communication resources, which is created by the feedback from MUs having no realistic chance of being scheduled for communication. These considerations motivate the use of selective feedback techniques in which only the MUs with channels good enough are allowed to feed back [29].

Having established the optimality of threshold feedback policies in the previous chapter, this chapter studies the important issue of how to set threshold values at MUs optimally to maximize downlink data rates without violating finite feedback constraints. To this end, we focus on the general setting of vector broadcast channels operating based on the concept of opportunistic beamforming [79]: Each MU calculates the SINR on each beam (random orthonormal beams in this context) and feeds back the SINR value and the index of the best beam to the base station. Upon retrieval of this information, the base station schedules the MU with the best SINR on each beam. Certainly, threshold feedback policies provide further feedback reductions for such systems, and therefore the following question is of practical and theoretical importance to answer: What is the optimal assignment of thresholds to MUs so that aggregate data rates over multiple beams are maximized subject to constraints on the average number of MUs feeding back per beam?

4.1.1 Related Works

On top of the work discussed in Chapter 3, related work also includes [65, 78]. In [65, 78], CSI parameters were quantized to reduce the feedback load for OBF. This approach cannot eliminate the linear feedback load growth alone, but it leads to further feedback reductions when combined with a user selection protocol. In this chapter, we solve the optimum threshold selection problem offline under statistical information about wireless channels. Once the thresholds are optimally assigned for user selection, it is an added design choice how to quantize SINR parameters, and the resulting performance analysis requires further investigation, which we do not address in this chapter.

Thresholding techniques are commonly seen in networks which do not utilize OBF as well [5,53,71,102,106]. In [106], the authors study a zero-force beamforming framework, and use an orthogonality threshold for selecting the MUs. By using the threshold, they are able to select a subset of MUs having channel vectors which are nearly orthogonal to each other. Optimum sumrate is achieved asymptotically. [5] also uses a zero-force beamforming framework, and a thresholding technique for user selection. In this technique, only the MUs with eigenvectors whose corresponding singular values are above a given threshold value are considered for selection. Among them, a subset of MUs having channel vectors which are nearly orthogonal to each other is selected. The authors provide sufficient and necessary conditions on the threshold value for the system to achieve optimum sum-rate capacity. These works are related in the sense that they use the same selective feedback technique, but different since they consider a framework which require perfect CSI at the base station.

In [71], the authors focused on exploiting multiuser diversity in a distributed manner for scalar multiple access channels by means of thresholds. Their MAC layer consisted of a collision channel model, and the thresholds were chosen to be the same for all MUs with identical channel statistics. Although we focus on the *dual* vector broadcast channels in this chapter without any attention on the multiple access feedback uplink, our results have some ramifications for the MAC problem studied in [71]. First of all, our homogenous threshold optimality results imply that using different threshold levels for different MUs with identical channel conditions may further improve the data rates reported in [71]. Secondly, they provide a cross-layer design parameter for the number of MUs to be multiplexed on the uplink (for feedback) without any noticeable performance degradation at the downlink.

4.1.2 Contributions

As shown in the previous chapter, the optimization problem we face is over the familiar finite dimensional Euclidean spaces, but it turns out that the objective sum-rate function is not necessarily convex as a function of users' threshold values. Thus, we resort to the theory of majorization [51], and solve the optimum threshold selection problem by

identifying an underlying *Schur-concave* structure in the sum-rate function. In particular, we obtain sufficient conditions for the Schur-concavity of the sum-rate, and therefore for the rate optimality of *homogenous* threshold feedback policies in which all MUs use the same threshold for their feedback decisions. These conditions are provided for general fading models under some mild conditions on the resulting SINR distribution, which are satisfied by most common fading models such as Rayleigh, Rician and Nakagami fading.

A naive but intuitive approach to maximize the total downlink communication rate for a network with identical MUs experiencing statistically the same channel conditions is to use a homogenous threshold feedback policy satisfying feedback constraints. Rather surprisingly, our results reveal that this intuition does not always work here. We provide a simple counterexample in which only a single beam is used for the downlink communication with two MUs in a Rayleigh fading environment. In the high signal-to-noise-ratio (SNR) regime, necessary conditions for the Schur-concavity of the sum-rate are violated, and it becomes strictly suboptimal to use the same threshold value to mediate users' feedback decisions. Indeed, we prefer one MU over the other one by assigning a small threshold for this MU to minimize the feedback outage event probability, *i.e.*, the probability that none of the MUs feeds back. On the other hand, we show that the sum-rate is a Schur-concave function when the SNR is low, and therefore the homogenous threshold feedback policy satisfying feedback constraints with equality is the optimum policy to maximize the sum-rate in the low SNR regime. To put it in another way, we trade the power gain (due to multiuser diversity) for the degrees-of-freedom gain (due to minimum outage communication) in the high SNR regime, whereas the degrees-of-freedom gain is traded for the power gain in the low SNR regime. An extensive numerical study utilizing our sufficient conditions is also performed to illustrate optimality and sub-optimality regions for the homogenous threshold feedback policies for fading models other than Rayleigh fading such as Rician and Nakagami fading.

On the more positive side, we show that the sum-rate is always a Schur-concave function for all values of SNR when *two or more* orthonormal beams are used to simultaneously communicate with multiple MUs located in a Rayleigh fading environment. In this case, the downlink communication becomes interference limited, rather than noise lim-

ited, due to inter-beam interference, and therefore the behavior of the optimum threshold feedback policy becomes unchanged for all SNR values: Use the same threshold for all MUs such that the feedback constraint is satisfied with equality. For this fading scenario, the difference between communication rates achieved with and without user selection is also illustrated. In particular, when the threshold values are optimally set for large MU populations, there is almost no rate loss if the average number of MUs feeding back per beam is around five. From a practical point of view, this signifies a significant reduction in the feedback load without noticeable performance loss, and provides an important cross-layer design parameter for the higher MAC layer for multiplexing MUs on the uplink to feed back.

The rest of this chapter is organized as follows. We first formulate the problem in Section 4.2, and provide an overview of our main results in Section 4.3 without any formal proofs. We then introduce some key concepts from the theory of majorization in Section 4.4. Then, formal proofs are supplied in Section 4.5 and in related appendices. We present an extensive numerical and simulation study to illustrate the applications of the results to familiar fading models along with various engineering and design insights in Section 4.6. Section 4.7 concludes the chapter.

4.2 Problem Formulation

In this chapter, we study the vector broadcast channel model given in Subsection 2.2.1, and we extend the problem formulated in Chapter 3, where we established the optimality of threshold feedback policies. Firstly, we will recall some important information, which will facilitate exposition in the rest of the chapter.

To recall, the base station communicates with n MUs through M different orthogonal beams simultaneously. The base station has N_t transmit antennas, and each MU is equipped with a single receive antenna. The beams are assumed to be statistically identical, and MUs experience statistically independent channel conditions. $\gamma_{i,m}$ is the SINR at beam m at MU i, and $\boldsymbol{\gamma}_i \in \mathbb{R}_+^M$ represents the SINR vector at MU i. The elements of $\boldsymbol{\gamma}_i$ are identically distributed with a common marginal distribution F. We will assume that

F is continuous, and has the density f with support $\mathbb{R}_+$. The M-by-n SINR matrix of the whole n-user communication system is denoted by $\mathbf{\Gamma} \in \mathbb{R}_+^{M\times n}$.

The MUs will feedback using a decentralized general threshold feedback policy, which has been defined in Chapter 3. We restate the definition in this chapter for the sake of completeness.

Definition 3.3 *We say*

$$\boldsymbol{\mathcal{T}} = (\mathcal{T}_1, \ldots, \mathcal{T}_n)^\top$$

is a general threshold feedback policy *(GTFP) if, for all $i \in \mathcal{N}$, there is a threshold τ_i such that $\mathcal{T}_i(\gamma_i)$ generates a feedback packet containing* SINR *values $\{\gamma_{k,i}\}_{k\in\mathcal{I}_i}$ if and only if $\gamma_{k,i} \geq \tau_i$ for all $k \in \mathcal{I}_i \subseteq \mathcal{M}$. We call it a* homogenous general threshold feedback policy *if all MUs use the same threshold τ,* i.e., *$\tau_i = \tau$ for all $i \in \mathcal{N}$.*

We only consider GTFPs in this chapter to avoid repetition (the definition of maximum threshold feedback policy (MTFP) is given in Definition 3.4, and the subtle differences between GTFPs and MTFPs are discussed in Section 3.5). We will also drop the term general for simplicity.

Given a threshold feedback policy $\boldsymbol{\mathcal{T}}$, we have a random set of MUs $\mathcal{G}_m(\boldsymbol{\mathcal{T}}(\mathbf{\Gamma}))$ requesting beam $m \in \mathcal{M}$. Then, the downlink ergodic sum-rate achieved under the threshold feedback policy $\boldsymbol{\mathcal{T}}$ is given by

$$R(\boldsymbol{\mathcal{T}}) = \mathsf{E}_{\mathbf{\Gamma}}\left[r(\boldsymbol{\mathcal{T}}, \mathbf{\Gamma})\right] = \mathsf{E}_{\mathbf{\Gamma}}\left[\sum_{m=1}^{M} \log\left(1 + \max_{i\in\mathcal{G}_m(\boldsymbol{\mathcal{T}}(\mathbf{\Gamma}))} \gamma_{m,i}\right)\right], \tag{4.1}$$

where $r(\boldsymbol{\mathcal{T}}, \mathbf{\Gamma})$ is the *instantaneous* sum-rate achieved under the threshold feedback policy $\boldsymbol{\mathcal{T}}$. $r^m(\boldsymbol{\mathcal{T}}, \mathbf{\Gamma})$ and $R^m(\boldsymbol{\mathcal{T}})$ denote the instantaneous sum-rate and the ergodic sum-rate on beam m, respectively. Note that

$$r^m(\boldsymbol{\mathcal{T}}, \mathbf{\Gamma}) = \log\left(1 + \max_{i\in\mathcal{G}_m(\boldsymbol{\mathcal{T}}(\mathbf{\Gamma}))} \gamma_{m,i}\right), \tag{4.2}$$

and

$$R^m(\boldsymbol{\mathcal{T}}) = \mathsf{E}_{\boldsymbol{\Gamma}}\left[r^m(\boldsymbol{\mathcal{T}}, \boldsymbol{\Gamma})\right]. \tag{4.3}$$

We will use $R(\boldsymbol{\mathcal{T}})$ as the performance measure of a given feedback policy along the rate dimension.

Given a threshold feedback policy $\boldsymbol{\mathcal{T}}$, we will use the average number of MUs feeding back per beam $\Lambda(\boldsymbol{\mathcal{T}})$ to measure the performance of $\boldsymbol{\mathcal{T}}$ along the feedback dimension. We have

$$\Lambda(\boldsymbol{\mathcal{T}}) = \sum_{i=1}^{n} \mathsf{Pr}\left\{\gamma_{1,i} \geq \tau_i\right\}.$$

Given a finite feedback system with a feedback constraint λ, the main optimization problem to be solved now is to determine the optimal threshold values to maximize the downlink throughput, which is formally written as

$$\begin{array}{ll} \underset{\boldsymbol{\tau} \in \mathbb{R}_+^n}{\text{maximize}} & R(\boldsymbol{\tau}) \\ \text{subject to} & \sum_{i=1}^{n} \mathsf{Pr}\left\{\gamma_{1,i} \geq \tau_i\right\} \leq \lambda \end{array}. \tag{4.4}$$

The optimization problem in (4.4), which we call *optimal threshold selection problem*, is still not easy to solve, even for a simple two-user system, due to the non-convex objective function, and a constraint set depending on the distribution of SINR values. The complexity of the problem further increases with increasing numbers of MUs due to the dimensionality growth. Therefore, it is not possible to solve the optimal threshold selection problem in its full generality for a general n-user system. However, we can still search for a structure in the sum-rate function to solve the optimal threshold selection problem, which is what we will do in the remainder of this chapter.

More specifically, we will search for sufficient conditions to be satisfied by SINR distributions so that the sum-rate becomes a Schur-concave function of feedback probabilities. Roughly speaking, a Schur-concave function increases when the dispersion among the components of its argument decreases, which implies a solution for the optimization

problem in (4.4) is a homogenous threshold feedback policy in which thresholds are set according to

$$\boldsymbol{\tau}^{\star} = \left(F^{-1}\left(1-\frac{\lambda}{n}\right), \cdots, F^{-1}\left(1-\frac{\lambda}{n}\right)\right)^{\top} \tag{4.5}$$

if the sum-rate is a Schur-concave function. We make this intuitive idea rigorous below.

4.3 Main Results

The main results of this section are stated in Theorems 4.1 and 4.2. In these theorems, we view the sum-rate as a function of feedback probabilities. This approach does not limit the generality of our results since SINR probability density function is already assumed to have $\mathbb{R}_+$ as its support, and therefore there is a one-to-one correspondence between feedback threshold values τ_i and the feedback probabilities p_i, *i.e.*, $\tau_i = F^{-1}(1-p_i)$ for all $i \in \mathcal{N}$. This assumption is satisfied for many commonly used practical fading models such as Rayleigh, Rician and Nakagami fading. Our theorems are as follows, and q in Theorem 4.1 is an auxiliary random variable.

Theorem 4.1. *The sum-rate $R(\boldsymbol{p})$ is a Schur-concave function if*

$$\log(1+\gamma)(\lambda-2q) + \int_{F^{-1}(1+q-\lambda)}^{F^{-1}(1-q)} \frac{F(x)}{1+x}dx - (\lambda-2q)\log\left(1+F^{-1}(1+q-\lambda)\right) \geq 0 \tag{4.6}$$

for all $\gamma \geq 0$, $\lambda \in [0,2]$ and $\max\{0,\lambda-1\} \leq q \leq \frac{\lambda}{2}$.

Theorem 4.2. *The sum-rate $R(\boldsymbol{p})$ is Schur-concave if f is bounded at zero, and has the derivative f' satisfying*

$$f'\left(F^{-1}(x)\right) \leq -\frac{f\left(F^{-1}(x)\right)}{1+F^{-1}(x)} \tag{4.7}$$

for all $x \in [0,1]$.

The proofs of Theorems 4.1 and 4.2 require introduction of new notation, and involve several cases to analyze separately. We also need some key results from the theory of

majorization [51] to prove these results. Therefore, we have relegated their proofs to the following sections and appendixes. We now briefly discuss their implications.

We first note that the sufficient condition for the Schur-concavity of the sum-rate given in (4.6) is stronger than the one given in (4.7) in the sense that (4.6) always holds whenever (4.7) holds, but not vice versa. This is formally established in Section 4.5. Furthermore, since the first term in (4.6) is always positive, an easier condition to check for the Schur-concavity of the sum-rate function is

$$\int_{F^{-1}(1+q-\lambda)}^{F^{-1}(1-q)} \frac{F(x)}{1+x} dx - (\lambda - 2q) \log\left(1 + F^{-1}(1+q-\lambda)\right) \geq 0 \tag{4.8}$$

for all $\lambda \in [0,2]$ and $\max\{0, \lambda - 1\} \leq q \leq \frac{\lambda}{2}$. Further, we can bound (4.8) from below to obtain another sufficient condition as

$$(1+q-\lambda) \log\left(1 + F^{-1}(1-q)\right) - (1-q) \log\left(1 + F^{-1}(1+q-\lambda)\right) \geq 0, \tag{4.9}$$

for all $\lambda \in [0,2]$ and $\max\{0, \lambda - 1\} \leq q \leq \frac{\lambda}{2}$. For a two-user system, (4.8) is also necessary, *i.e.,* see Lemma 4.7 and discussions therein.

Although the conditions (4.8) and (4.9) are easy to verify numerically, they may not be tractable analytically. The integral expression in (4.8) is hard to evaluate in closed-form. Analytical verification of (4.9) is also difficult due to the presence of conflicting forces working in opposite directions to increase/decrease the value of the bound. For example, the pre-log factor of the first term in (4.9), which is $1+q-\lambda$, is smaller than the pre-log factor of the second term, which is $1-q$, for $\max\{0, \lambda - 1\} \leq q \leq \frac{\lambda}{2}$. Conversely, for $\max\{0, \lambda - 1\} \leq q \leq \frac{\lambda}{2}$, $F^{-1}(1-q)$ appearing inside the logarithm in the first term is greater than $F^{-1}(1+q-\lambda)$ appearing inside the logarithm in the second term.

On the other hand, the sufficient condition for the Schur-concavity of the sum-rate function given in Theorem 4.2 turns out to be much easier to deal with analytically although it looks more complex than (4.8) and (4.9). In particular, it provides an *almost* complete characterization for the solution of optimal threshold selection problem for richly scattered Rayleigh fading environments. More precisely, (4.7) is always satisfied for all values of ρ for Rayleigh fading channels whenever $M \geq 2$. Hence, the sum-rate is al-

ways a Schur-concave function of feedback probabilities in this case, and is maximized if thresholds are chosen according to (4.5). In Section 4.6, we provide a detailed discussion for the optimality and sub-optimality of homogenous threshold feedback policies for Rayleigh fading channels as well as other wireless channel models. Next, we will briefly introduce some key concepts from the theory of majorization to be used later in our analysis.

4.4 Majorization

For a vector $\boldsymbol{x}$ in $\mathbb{R}^n$, we denote its ordered coordinates by

$$x_{(1)} \geq \cdots x_{(k-1)} \geq x_{(k)} \geq x_{(k+1)} \geq \cdots x_{(n)}$$

For $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$, we say $\boldsymbol{x}$ *majorizes* $\boldsymbol{y}$ and write it as $\boldsymbol{x} \succeq_{\mathrm{M}} \boldsymbol{y}$ if we have

$$\sum_{i=1}^{k} x_{(i)} \geq \sum_{i=1}^{k} y_{(i)}$$

when $k = 1, \ldots, n-1$, and

$$\sum_{i=1}^{n} x_{(i)} = \sum_{i=1}^{n} y_{(i)}.$$

A function $\varphi : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be *Schur-convex* if $\boldsymbol{x} \succeq_{\mathrm{M}} \boldsymbol{y}$ implies $\varphi(\boldsymbol{x}) \geq \varphi(\boldsymbol{y})$, and φ is *Schur-concave* if $-\varphi$ is Schur-convex. Schur-convex/concave functions often arise in mathematical analysis and engineering applications [3, 95]. For example, every function that is concave (convex) and symmetric is also a Schur-concave (Schur-convex) function.

A Schur-concave function tends to increase when the components of its argument become more similar. We will establish conditions under which the sum-rate becomes a Schur-concave function, which will, in turn, imply the optimality of homogenous threshold feedback policies. The following lemma is helpful in establishing these conditions.

Lemma 4.1. *Let φ be a real-valued function defined on $\mathbb{R}_+^n$, and*

$$\mathcal{D} = \{\boldsymbol{z} \in \mathbb{R}_+^n : z_1 \geq \cdots \geq z_n\}.$$

Then, φ is a Schur-convex function if and only if, for all $\boldsymbol{z} \in \mathcal{D}$ and $i = 1, \ldots, n-1$,

$$\varphi\left(z_1, \ldots, z_{i-1}, z_i + \epsilon, z_{i+1} - \epsilon, z_{i+2}, \ldots, z_n\right)$$

is increasing in ϵ over the region $0 \leq \epsilon \leq \min\{z_{i-1} - z_i, z_{i+1} - z_{i+2}\}$. [1]

It can be seen that the coordinates z_i and z_{i+1} are systematically altered by using the parameter ϵ, and the constraints on ϵ eliminate any violation in the order. Interested readers are referred to [51] for more insights on the theory of majorization. Now, we will see how we can use this theory to identify the Schur-concave structure in the objective rate function.

4.5 Schur-concavity Analysis for the Sum-rate

The main objective is to establish sufficient conditions on the SINR distributions for the Schur-concavity of the sum-rate function. Again, we focus on the first beam to explain our proof ideas without any loss of generality since all beams are statistically identical. We start by analyzing the sum-rate as a function of thresholds as given in (4.4) to establish three important lemmas. Next, we will incorporate the feedback constraint into our optimization problem by interpreting the sum-rate as a function of feedback probabilities. Using these results, we will finally establish the underlying Schur-concave structure in the sum-rate function through the theory of majorization.

[1] At the end points $i = 1$, $i = n - 1$, the condition is modified accordingly.

4.5.1 Rate as a Function of Thresholds

Consider thresholds in increasing order, *i.e.*, $\tau_{\pi(1)} \leq \cdots \leq \tau_{\pi(i)} \leq \tau_{\pi(i+1)} \leq \cdots \leq \tau_{\pi(n)}$. Based on Lemma 4.1, it is enough to consider

$$R^1\left(\tau_{\pi(i+1)}+\epsilon, \tau_{\pi(i)}-\epsilon\right) = R^1\left(\tau_{\pi(n)}, \ldots, \tau_{\pi(i+1)}+\epsilon, \tau_{\pi(i)}-\epsilon, \ldots, \tau_{\pi(1)}\right),$$

to identify the underlying Schur-concave structure in the sum-rate function.[2] However, analysis of this function is still complex. Therefore, we resort to the following divide-and-conquer approach.

With a slight abuse of notation, we define the truncated SINR on beam m at MU i as

$$\bar{\gamma}_{m,i} = \gamma_{m,i} 1_{\{\gamma_{m,i} \geq \tau_i\}}.$$

Let

$$\mathcal{N}' = \{k \in \mathcal{N} : k \neq \pi(i) \ \& \ \pi(i+1)\}.$$

Also, let

$$\bar{\gamma}^{\star}_{\mathcal{N}'} = \max_{k \in \mathcal{N}'} \bar{\gamma}_{1,k},$$

which is the maximum truncated SINR on beam 1 among the MUs in $\mathcal{N}'$. The instantaneous rate on beam 1 as a function of $\bar{\gamma}_{1,\pi(i)}$, $\bar{\gamma}_{1,\pi(i+1)}$ and $\bar{\gamma}^{\star}_{\mathcal{N}'}$ is

$$r^1\left(\bar{\gamma}_{1,\pi(i+1)}, \bar{\gamma}_{1,\pi(i)}, \bar{\gamma}^{\star}_{\mathcal{N}'}\right) = \log\left(1 + \max\left\{\bar{\gamma}_{1,\pi(i+1)}, \bar{\gamma}_{1,\pi(i)}, \bar{\gamma}^{\star}_{\mathcal{N}'}\right\}\right). \tag{4.10}$$

We fix the thresholds and the SINR values of all MUs in $\mathcal{N}'$.[3] Randomness is now associated only with MUs $\pi(i)$ and $\pi(i+1)$. Therefore,

$$R^1\left(\tau_{\pi(i+1)}, \tau_{\pi(i)} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right) = \mathsf{E}\left[r^1\left(\bar{\gamma}_{1,\pi(i+1)}, \bar{\gamma}_{1,\pi(i)}, \bar{\gamma}^{\star}_{\mathcal{N}'}\right) | \bar{\gamma}^{\star}_{\mathcal{N}'}\right]. \tag{4.11}$$

[2] We suppress the dependency of R^1 on $\tau_{\pi(k)}, k \neq i, i+1$ here and later in the chapter when we focus only on thresholds $\tau_{\pi(i)}$ and $\tau_{\pi(i+1)}$.

[3] Fixing random SINR values means conditioning on them in the probabilistic sense. Out of the truncated SINR values which are being fixed, only $\bar{\gamma}^{\star}_{\mathcal{N}'}$ will affect the rate expression, because the beam will be allocated to the MU with the maximum truncated SINR. Therefore, it is sufficient just to condition on $\bar{\gamma}^{\star}_{\mathcal{N}'}$.

As shown later in the chapter, this approach helps us to use the results derived for a two-user system to simplify our analysis. Therefore, considering a two-user system first, the rate on beam 1 as a function of the thresholds is explicitly given in Lemma 4.2.

Lemma 4.2. *The rate on beam* 1 *of a two-user system is equal to*

$$R^1(\boldsymbol{\tau}) = \int_{\tau_{\pi(2)}}^{\infty} \log(1+x) dF^2(x) + F\left(\tau_{\pi(2)}\right) \int_{\tau_{\pi(1)}}^{\tau_{\pi(2)}} \log(1+x) dF(x).$$

Proof. See Appendix 4.8.1. □

Coming back to the general n-user scenario, it is not tractable to obtain the rate explicitly as we have done in the previous lemma. However, we can explicitly write down an expression for $R^1\left(\tau_{\pi(i+1)}, \tau_{\pi(i)} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right)$. $R^1\left(\tau_{\pi(i+1)}, \tau_{\pi(i)} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right)$ is parameterized by $\bar{\gamma}^{\star}_{\mathcal{N}'}$, and its shape depends on the value of $\bar{\gamma}^{\star}_{\mathcal{N}'}$. Three cases of interest are $\bar{\gamma}^{\star}_{\mathcal{N}'} > \tau_{\pi(i+1)}$, $\bar{\gamma}^{\star}_{\mathcal{N}'} < \tau_{\pi(i)}$ and $\tau_{\pi(i)} \leq \bar{\gamma}^{\star}_{\mathcal{N}'} \leq \tau_{\pi(i+1)}$. We will now establish three important lemmas for these three cases, which will be useful in interpreting the rate function. The two-user rate expression given in Lemma 4.2 functions as a building block to obtain beam 1 rate expressions in these cases. We will start with the case $\bar{\gamma}^{\star}_{\mathcal{N}'} > \tau_{\pi(i+1)}$.

Lemma 4.3. *If* $\bar{\gamma}^{\star}_{\mathcal{N}'} > \tau_{\pi(i+1)}$, $R^1\left(\tau_{\pi(i+1)}, \tau_{\pi(i)} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right)$ *is given by*

$$R_0\left(\bar{\gamma}^{\star}_{\mathcal{N}'}\right) = \mathsf{Pr}\left\{\xi^{\star}_{i+1,i} \leq \bar{\gamma}^{\star}_{\mathcal{N}'} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right\} \log\left(1+\bar{\gamma}^{\star}_{\mathcal{N}'}\right) + \mathsf{E}\left[\log\left(1+\xi^{\star}_{i+1,i}\right) 1_{\left\{\xi^{\star}_{i+1,i} > \bar{\gamma}^{\star}_{\mathcal{N}'}\right\}} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right],$$

where $\xi^{\star}_{i+1,i} = \max\left\{\gamma_{1,\pi(i+1)}, \gamma_{1,\pi(i)}\right\}$.

Proof. See Appendix 4.8.2. □

Note that $R^1\left(\tau_{\pi(i+1)}, \tau_{\pi(i)} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right)$ depends only on $\bar{\gamma}^{\star}_{\mathcal{N}'}$ but not on $\tau_{\pi(i)}$ and $\tau_{\pi(i+1)}$ when $\bar{\gamma}^{\star}_{\mathcal{N}'} > \tau_{\pi(i+1)}$. The next lemma provides an analogous expression for $R^1\left(\tau_{\pi(i+1)}, \tau_{\pi(i)} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right)$ when $\bar{\gamma}^{\star}_{\mathcal{N}'} < \tau_{\pi(i)}$.

Lemma 4.4. *If* $\bar{\gamma}^{\star}_{\mathcal{N}'} < \tau_{\pi(i)}$, $R^1\left(\tau_{\pi(i+1)}, \tau_{\pi(i)} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right)$ *is given by*

$$R_1\left(\tau_{\pi(i+1)}, \tau_{\pi(i)} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right) = \int_{\tau_{\pi(i+1)}}^{\infty} \log(1+x) dF^2(x) + F\left(\tau_{\pi(i+1)}\right) \int_{\tau_{\pi(i)}}^{\tau_{\pi(i+1)}} \log(1+x) dF(x) + \log\left(1+\bar{\gamma}^{\star}_{\mathcal{N}'}\right) F\left(\tau_{\pi(i)}\right) F\left(\tau_{\pi(i+1)}\right).$$

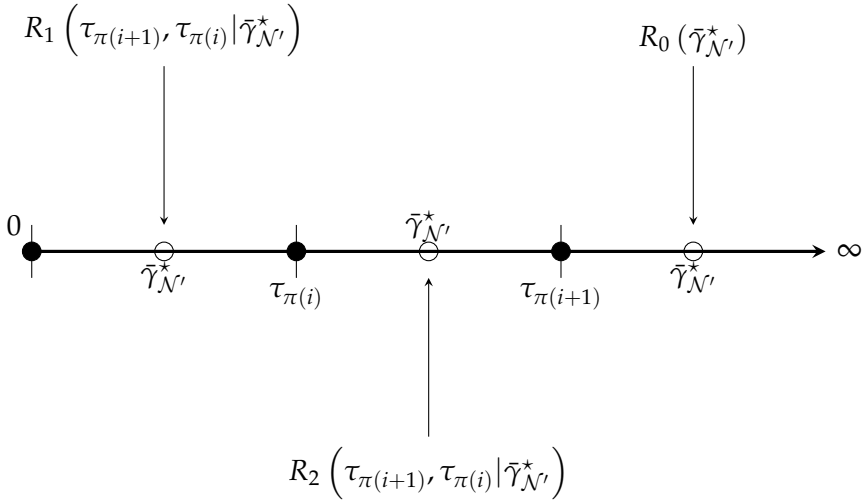


Figure 4.1: Beam 1 rate as a function of thresholds for different values of $\bar{\gamma}^{\star}_{\mathcal{N}'}$

Proof. See Appendix 4.8.2. □

Finally, we look at the case where $\tau_{\pi(i)} \leq \bar{\gamma}^{\star}_{\mathcal{N}'} \leq \tau_{\pi(i+1)}$.

Lemma 4.5. *If* $\tau_{\pi(i)} \leq \bar{\gamma}^{\star}_{\mathcal{N}'} \leq \tau_{\pi(i+1)}$, $R^1\left(\tau_{\pi(i+1)}, \tau_{\pi(i)} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right)$ *is given by*

$$R_2\left(\tau_{\pi(i+1)}, \tau_{\pi(i)} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right) = \int_{\tau_{\pi(i+1)}}^{\infty} \log(1+x) dF^2(x) + F\left(\tau_{\pi(i+1)}\right) \int_{\bar{\gamma}^{\star}_{\mathcal{N}'}}^{\tau_{\pi(i+1)}} \log(1+x) dF(x) + \log\left(1+\bar{\gamma}^{\star}_{\mathcal{N}'}\right) F\left(\tau_{\pi(i+1)}\right) F\left(\bar{\gamma}^{\star}_{\mathcal{N}'}\right).$$

Proof. See Appendix 4.8.2. □

For the final two cases, we note that $R^1\left(\tau_{\pi(i+1)}, \tau_{\pi(i)} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right)$ depends both on threshold values $\tau_{\pi(i)}$ and $\tau_{\pi(i+1)}$, and on $\bar{\gamma}^{\star}_{\mathcal{N}'}$. The results of these three lemmas have been graphically summarized in Fig. 4.1.

If $\bar{\gamma}^{\star}_{\mathcal{N}'} = \tau_{\pi(i)}$, R_1 and R_2 in Lemmas 4.4 and 4.5 evaluate to the same expression. Similarly, if $\bar{\gamma}^{\star}_{\mathcal{N}'} = \tau_{\pi(i+1)}$, R_0 and R_2 in Lemmas 4.3 and 4.5 evaluate to the same expression. This shows that the rate as a function of $\bar{\gamma}^{\star}_{\mathcal{N}'}$ is continuous at $\tau_{\pi(i)}$ and $\tau_{\pi(i+1)}$.

Given the initial threshold values $\left\{\tau_{\pi(k)}\right\}_{k=1}^{n}$, the first step to discover the Schur-

concave structure in the sum-rate function is to analyze the behavior of the function

$$g_T(\epsilon) = R^1\left(\tau_{\pi(i+1)} + \epsilon, \tau_{\pi(i)} - \epsilon | \bar{\gamma}^\star_{\mathcal{N}'}\right)$$

for

$$\epsilon \in \left[0, \min\left\{\tau_{\pi(i)} - \tau_{\pi(i-1)}, \tau_{\pi(i+2)} - \tau_{\pi(i+1)}\right\}\right]$$

by making use of Lemma 4.1. This is now a scalar problem. At this point, it is more useful to interpret the sum-rate as a function of feedback probabilities since the feedback constraint in (4.4) is in terms of these probabilities. This interpretation helps us to incorporate the feedback constraints into our optimization problem more easily, as will be shown next.

4.5.2 Rate as a Function of Feedback Probabilities

There is a one-to-one correspondence between feedback thresholds $\tau_{\pi(i)}$ and feedback probabilities $p_{\pi(i)}$ since f has the support $\mathbb{R}_+$, *i.e.*, $\tau_{\pi(i)} = F^{-1}(1 - p_{\pi(i)})$. Hence, we can represent $R^1\left(\tau_{\pi(i+1)}, \tau_{\pi(i)} | \bar{\gamma}^\star_{\mathcal{N}'}\right)$ as $R^1\left(p_{\pi(i)}, p_{\pi(i+1)} | \bar{\gamma}^\star_{\mathcal{N}'}\right)$ without any ambiguity. With this interpretation, the optimization problem in (4.4) can be considered as the problem of finding optimum feedback probability vector

$$\boldsymbol{p}^\star = (p_1^\star, \ldots, p_n^\star)^\top$$

in $[0,1]^n$ subject to the feedback constraint

$$\sum_{i=1}^{n} p_i^\star \leq \lambda,$$

where $p_i = \mathsf{Pr}\{\gamma_{1,i} \geq \tau_i\}$. Indeed, it is easy to see that any feedback policy solving (4.4) must achieve the feedback constraint with equality, *i.e.*,

$$\sum_{i=1}^{n} p_i^\star = \lambda.$$

Since F is monotone increasing, we have

$$p_{\pi(1)} \geq p_{\pi(2)} \geq \cdots \geq p_{\pi(i)} \geq p_{\pi(i+1)} \geq \cdots \geq p_{\pi(n)}.$$

Focusing on $p_{\pi(i)}$ and $p_{\pi(i+1)}$, we have the feedback level

$$\lambda_i = p_{\pi(i)} + p_{\pi(i+1)},$$

and other probabilities give us natural boundaries on $p_{\pi(i)}$ and $p_{\pi(i+1)}$ as such $p_{\pi(i+2)} \leq p_{\pi(i+1)} \leq p_{\pi(i)} \leq p_{\pi(i-1)}$. Without violating these boundaries, we will vary $p_{\pi(i)}$ and $p_{\pi(i+1)}$ by keeping λ_i constant.

Similar to the previous part, we start our analysis by focusing on a two-user system. Given a feedback constraint $\lambda > 0$, we can restrict our search for the optimal feedback probability vector to the plane given by

$$p_{\pi(1)} + p_{\pi(2)} = \lambda.$$

On this plane, we write the rate function $R^1(\boldsymbol{p})$ as a function of only $p_{\pi(2)}$ without any ambiguity. The communication rate on this plane as a function of $p_{\pi(2)}$ is given below.

Lemma 4.6. *The rate on beam* 1 *of a two-user system on the plane*

$$\mathcal{P} = \left\{ \boldsymbol{p} \in [0,1]^2 : p_{\pi(1)} + p_{\pi(2)} = \lambda \right\}$$

as a function of $p_{\pi(2)}$ *is equal to*

$$R^1\left(p_{\pi(2)}\right) = \int_{F^{-1}\left(1-p_{\pi(2)}\right)}^{\infty} \log(1+x)\, dF^2(x) + \left(1 - p_{\pi(2)}\right) \int_{F^{-1}\left(1+p_{\pi(2)}-\lambda\right)}^{F^{-1}\left(1-p_{\pi(2)}\right)} \log(1+x)\, dF(x).$$

for $\max\{0, \lambda - 1\} \leq p_{\pi(2)} \leq \frac{\lambda}{2}$

Proof. Follows from a direct substitution of $\tau_{\pi(2)} = F^{-1}\left(1 - p_{\pi(2)}\right)$ in Lemma 4.2. □

F^{-1} in the expression above represents the functional inverse of F. We give the first derivative of the two-user rate in the following lemma.

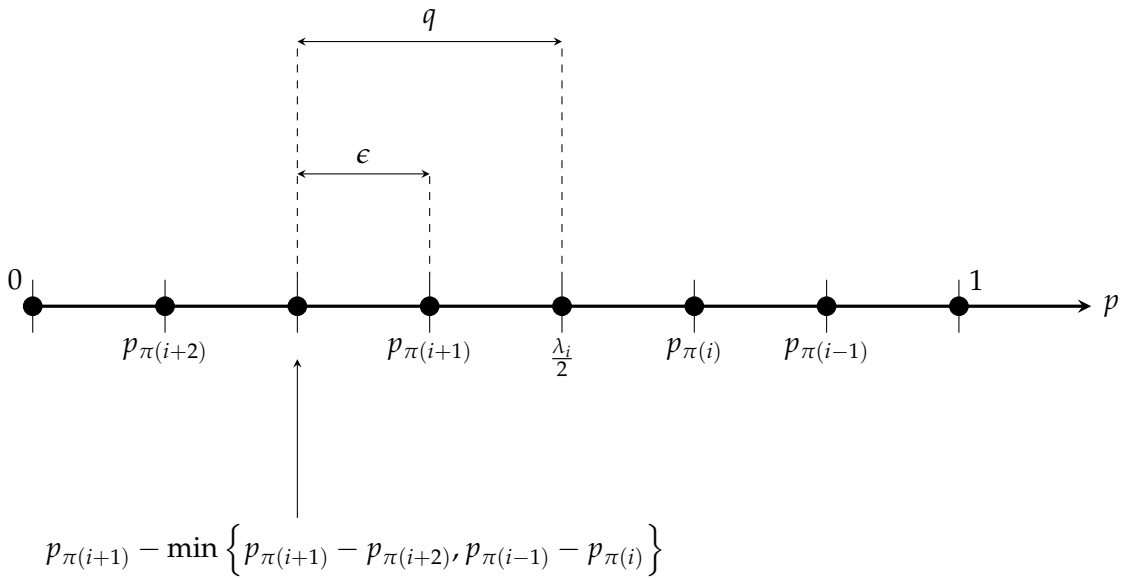


Figure 4.2: Ordered feedback probabilities, and the range of q and ϵ.

Lemma 4.7. *The first derivative of* $R^1\left(p_{\pi(2)}\right)$ *on*

$$\mathcal{P} = \left\{\boldsymbol{p} \in [0,1]^2 : p_{\pi(1)} + p_{\pi(2)} = \lambda\right\}$$

is equal to

$$\frac{dR^1\left(p_{\pi(2)}\right)}{dp_{\pi(2)}} = \int_{F^{-1}\left(1+p_{\pi(2)}-\lambda\right)}^{F^{-1}\left(1-p_{\pi(2)}\right)} \frac{F(x)}{1+x}dx - \left(\lambda - 2p_{\pi(2)}\right)\log\left(1 + F^{-1}\left(1+p_{\pi(2)}-\lambda\right)\right). \quad (4.12)$$

for $\max\{0, \lambda - 1\} \leq p_{\pi(2)} \leq \frac{\lambda}{2}$.

Proof. Follows directly after differentiating the rate expression in Lemma 4.6. □

We note that Lemma 4.1 implies the necessity of $\frac{dR^1\left(p_{\pi(2)}\right)}{dp_{\pi(2)}} \geq 0$ for all

$$p_{\pi(2)} \in \left[\max\{0, \lambda - 1\}, \frac{\lambda}{2}\right]$$

and $\lambda \in [0,2]$ for the Schur-concavity of the two-user sum-rate. Consider now the n-

user scenario. Given the initial feedback probabilities $\left\{p_{\pi(k)}\right\}_{k=1}^{n}$, we need to analyze the behavior of the function

$$g_p(\epsilon) = R^1\left(p_{\pi(i)} + \epsilon, p_{\pi(i+1)} - \epsilon | \bar{\gamma}^{\star}_{\mathcal{N}'}\right) \tag{4.13}$$

for

$$\epsilon \in \left[0, \min\left\{p_{\pi(i-1)} - p_{\pi(i)}, p_{\pi(i+1)} - p_{\pi(i+2)}\right\}\right]$$

to discover Schur-concavity of the rate function by Lemma 4.1. We have already discussed how we can vary $p_{\pi(2)}$ by keeping λ constant for the two-user case. Analysis of the general n-user scenario is not fundamentally different from the two-user scenario, and a similar technique used for the analysis of the two-user rate function can still be applied for the general n-user case without violating the boundary conditions on feedback probabilities. That is, we introduce an auxiliary variable $q \in \mathcal{P}_{i+1}$, replace $p_{\pi(i+1)} - \epsilon$ with q and $p_{\pi(i)} + \epsilon$ with $\lambda_i - q$, and write $R^1\left(p_{\pi(i)} + \epsilon, p_{\pi(i+1)} - \epsilon | \bar{\gamma}^{\star}_{\mathcal{N}'}\right)$ as a function of q, where

$$\mathcal{P}_{i+1} = \left[p_{\pi(i+1)} - \min\left\{p_{\pi(i+1)} - p_{\pi(i+2)}, p_{\pi(i-1)} - p_{\pi(i)}\right\}, \frac{\lambda_i}{2}\right].$$

Fig. 4.2 provides a graphical representation for the selection of q. By using Lemma 4.3, 4.4 and 4.5, we have

$$\begin{aligned} R^1\left(q|\bar{\gamma}^{\star}_{\mathcal{N}'}\right) = R_0\left(\bar{\gamma}^{\star}_{\mathcal{N}'}\right) 1_{\left\{q>1-F\left(\bar{\gamma}^{\star}_{\mathcal{N}'}\right)\right\}} + R_1\left(q|\bar{\gamma}^{\star}_{\mathcal{N}'}\right) 1_{\left\{q>\lambda_i-\left(1-F\left(\bar{\gamma}^{\star}_{\mathcal{N}'}\right)\right)\right\}} \\ + R_2\left(q|\bar{\gamma}^{\star}_{\mathcal{N}'}\right) 1_{\left\{q\leq 1-F\left(\bar{\gamma}^{\star}_{\mathcal{N}'}\right) \,\&\, q\leq\lambda_i-\left(1-F\left(\bar{\gamma}^{\star}_{\mathcal{N}'}\right)\right)\right\}} \end{aligned} \tag{4.14}$$

for $q \in \mathcal{P}_{i+1}$.

Some insights about (4.14) are as follows. Let

$$q_{\min} = p_{\pi(i+1)} - \min\left\{p_{\pi(i+1)} - p_{\pi(i+2)}, p_{\pi(i-1)} - p_{\pi(i)}\right\},$$

and assume $1 - F\left(\bar{\gamma}^{\star}_{\mathcal{N}'}\right) \leq \frac{\lambda_i}{2}$. If $1 - F\left(\bar{\gamma}^{\star}_{\mathcal{N}'}\right) < q_{\min}$, $R^1\left(q|\bar{\gamma}^{\star}_{\mathcal{N}'}\right)$ is equal to $R_0\left(\bar{\gamma}^{\star}_{\mathcal{N}'}\right)$ for all $q \in \left[q_{\min}, \frac{\lambda_i}{2}\right]$. On the other hand, if $1 - F\left(\bar{\gamma}^{\star}_{\mathcal{N}'}\right) \geq q_{\min}$, $R^1\left(q|\bar{\gamma}^{\star}_{\mathcal{N}'}\right)$ first becomes equal

to $R_2\left(q|\bar{\gamma}^{\star}_{\mathcal{N}'}\right)$ and then equal to $R_0\left(\bar{\gamma}^{\star}_{\mathcal{N}'}\right)$ as q changes from $q_{\min}$ to $\frac{\lambda_i}{2}$. This behavior is graphically depicted in Fig. 4.3.[4] Therefore, the rate in this case can be visualized as a concatenation of two functions with a gluing point at $1-F\left(\bar{\gamma}^{\star}_{\mathcal{N}'}\right)$. Similar explanations can be given for $1-F\left(\bar{\gamma}^{\star}_{\mathcal{N}'}\right) > \frac{\lambda_i}{2}$.

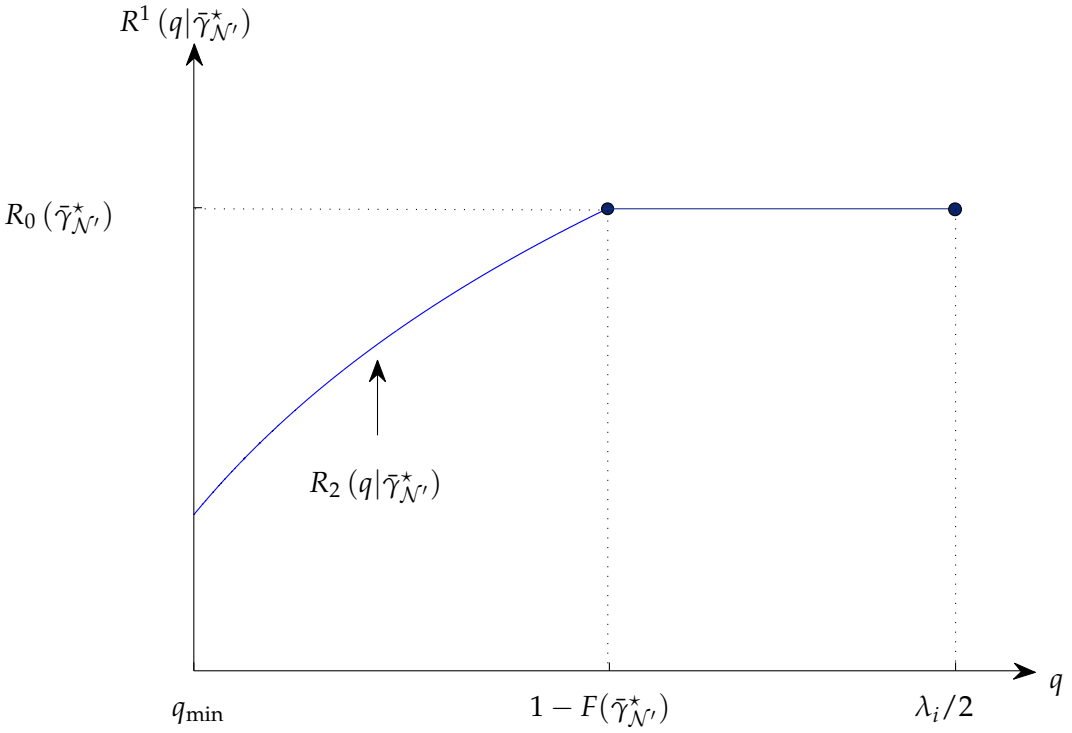


Figure 4.3: A pictorial representation for the rate expression in (4.14) for $q_{\min} \leq 1-F\left(\bar{\gamma}^{\star}_{\mathcal{N}'}\right) \leq \frac{\lambda_i}{2}$.

4.5.3 Schur-concavity of the Sum-rate Function

Building upon our analysis above, we will obtain sufficient conditions for the Schur-concavity of the sum-rate in this part. We start our analysis by first providing a proof for Theorem 4.1. We restated Theorem 4.1 below for the sake of completeness.

[4] The plot may not be exactly accurate. It is just given to conceptualize the behavior of the rate function.

Theorem 4.1: The sum-rate $R(\boldsymbol{p})$ is a Schur-concave function if

$$\log(1+\gamma)(\lambda-2q)+\int_{F^{-1}(1+q-\lambda)}^{F^{-1}(1-q)}\frac{F(x)}{1+x}dx-(\lambda-2q)\log\left(1+F^{-1}(1+q-\lambda)\right)\geq 0$$

for all $\gamma \geq 0$, $\lambda \in [0,2]$ and $\max\{0,\lambda-1\} \leq q \leq \frac{\lambda}{2}$.

Proof. Refer Appendix 4.8.3 for the proof. □

Second, we provide a proof for Theorem 4.2 based on Theorem 4.1. The new sufficient condition for the Schur-concavity of the sum-rate function is obtained by means of a second order analysis. Although complex looking, it turns out to be much easier to deal with analytically as illustrated for Rayleigh fading channels in the next section. Again, we restate Theorem 4.2 below for the sake of completeness.

Theorem 4.2: The sum-rate $R(\boldsymbol{p})$ is Schur-concave if f is bounded at zero, and has the derivative f' satisfying

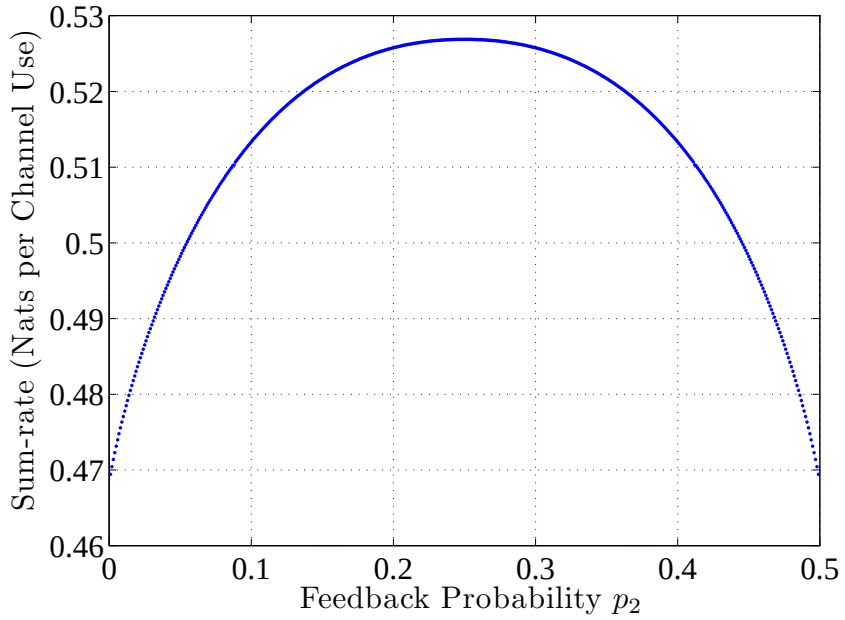
$$f'\left(F^{-1}(x)\right)\leq-\frac{f\left(F^{-1}(x)\right)}{1+F^{-1}(x)}$$

for all $x \in [0,1]$.

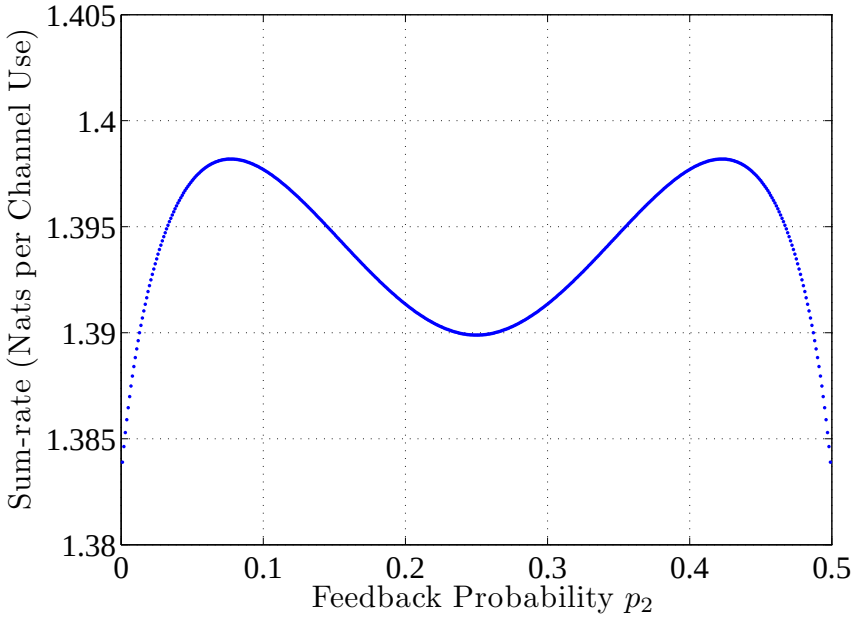
Proof. Refer Appendix 4.8.4 for the proof. □

4.6 Applications and Discussion

In this section, we will apply our results derived in Section 4.5 to well known fading channel models. We will also discuss the intuition behind the resulting performance figures. We start our discussion with Rayleigh fading channels, which is one of the most commonly used channel models in the literature, *e.g.*, see [9,82,83], and closely approximates measured data rates in densely populated urban areas [14].



(a) Behavior of the sum-rate as a function of the feedback probability p_2 of the second MU for the first example. ($M = 1, \lambda = 0.5$ and $\rho = 0$ [dB])



(b) Behavior of the sum-rate as a function of the feedback probability p_2 of the second MU for the second example. ($M = 1, \lambda = 0.5$ and $\rho = 10$ [dB])

Figure 4.4: Numerical examples illustrating the optimality and sub-optimality of homogenous threshold policies for different network configurations.

4.6.1 Rayleigh Fading Channels

Consider the Rayleigh fading channel model in which $h_{k,i}, k = 1, \ldots, N_t$ and $i = 1, \ldots, n$, are assumed to be i.i.d. with the common distribution $\mathcal{CN}(0,1)$, where $\mathcal{CN}(\mu, \sigma^2)$ represents the *circularly-symmetric complex Gaussian* distribution with mean μ and variance σ^2. Recall that the background noise is the unit power (complex) Gaussian noise, and therefore ρ is interpreted as the average SNR below.

For this channel model, the SINR distribution function F and the associated probability density function f can be given as

$$F(x) = 1 - \frac{e^{-\frac{x}{\rho}}}{(x+1)^{M-1}} \tag{4.15}$$

and

$$f(x) = \frac{e^{-\frac{x}{\rho}}}{(x+1)^{M}} \left[\frac{1}{\rho}(x+1) + M - 1 \right], \tag{4.16}$$

respectively [79]. More details on obtaining these expressions can be found in Appendix 5.9.1 of Chapter 5. An important quantity of interest to apply our results in Theorems 4.1 and 4.2 is the functional inverse, F^{-1}, of F. The next lemma provides an analytical expression for F^{-1} for Rayleigh fading channels.

Lemma 4.8. *F^{-1} is equal to*

$$F^{-1}(x) = \begin{cases} -1 + (M-1)\rho W\left(\frac{\exp\left(\frac{1}{(M-1)\rho}\right)}{(M-1)\rho} (1-x)^{\frac{1}{1-M}} \right) & \text{if } M \geq 2 \\ -\rho \log(1-x) & \text{if } M = 1 \end{cases},$$

where $x \in [0,1]$ and W is the Lambert W function given by the defining equation

$$W(x)\exp(W(x)) = x$$

for $x \geq -\frac{1}{e}$.

Proof. See Appendix 4.8.5. □

To motivate the discussion below, we start by providing two simple numerical examples; first of which illustrates a network configuration in which homogenous threshold feedback policies are optimal, whereas the second example provides another network configuration in which homogenous threshold feedback policies are strictly suboptimal.

Consider two MUs located in a Rayleigh fading environment, *i.e.,* all channel (amplitude) gains are random with distribution $\mathcal{CN}(0,1)$. M and λ are chosen to be $M = 1$ and $\lambda = 0.5$ in both examples below. We set ρ to 0 [dB] in the first example, while it is set to 10 [dB] in the second one. Since all MUs are identical with identical fading characteristics in this set-up, it is intuitively expected that a homogenous threshold feedback policy must be optimal, and solve the rate maximization problem in (4.4) under both network configurations.

This is indeed correct for the first network configuration as shown in Fig 4.4(a). The sum-rate is clearly maximized at $\boldsymbol{p} = (0.25, 0.25)^\top$, and therefore the homogenous threshold feedback policy with thresholds set as $\boldsymbol{\tau}_{\text{homo}} = (\log(4), \log(4))^\top$ solves (4.4). However, this intuition does not always work as illustrated by the second example. In this case, the homogenous threshold feedback policy equalizing the feedback probabilities of MUs becomes strictly suboptimal, *i.e.,* see Fig. 4.4(b). This shows that $R(\boldsymbol{p})$ is not a Schur-concave function of feedback probabilities for these selections of model parameters, and hence, it is not necessarily maximized at $\boldsymbol{p} = (0.25, 0.25)^\top$. We note that the selection of parameters in both examples is just for elucidatory purposes, and the same arguments continue to hold for other values of λ.

This discussion motivates the following question: When are homogenous threshold feedback policies optimal for Rayleigh fading channels? The answer is supplied by the following two theorems.

Theorem 4.3. *For Rayleigh fading environments with $M = 1$ and $\rho \leq 1$, $R(\boldsymbol{p})$ is a Schur-concave function of feedback probabilities, and therefore the homogenous threshold feedback policy satisfying feedback constraints with equality solves* (4.4) *when $M = 1$ and $\rho \leq 1$.*

Theorem 4.4. *For Rayleigh fading environments with $M > 1$, $R(\boldsymbol{p})$ is a Schur-concave function of feedback probabilities, and therefore the homogenous threshold feedback policy satisfying feedback constraints with equality solves* (4.4) *when $M > 1$.*

Proof. See Appendix 4.8.6. □

Since the proofs are similar, and are based on the sufficient condition established in Theorem 4.2, we skip the proof of Theorem 4.3 to avoid repetitions. Theorem 4.3 shows that it is enough to have ρ smaller than or equal to 1 to ensure the optimality of homogenous threshold feedback policies for Rayleigh fading environments when only a single beam is used for the downlink communication. Since F in (4.15) does not depend on N_t, the same result continues to hold for $N_t > 1$ as long as multiple transmit antennas are used to form a single beam as in [97].

On the other hand, Theorem 4.4 provides an extension of Theorem 4.3 to multiple beams. Theorem 4.4 is promising for multiuser MIMO downlink communication in a Rayleigh fading environment because it shows that homogenous threshold feedback policies are always optimal if multiple beams are used to communicate with multiple MUs simultaneously. Although the optimality of homogenous threshold feedback policies strongly depends on the properties of the underlying fading process modulating received signal strengths and the background noise level present in the system for the single beam case, this is not true anymore for multiple beams. More intuition is provided on this point later.

From a theoretical viewpoint, it is surprising to see that a property holding in the setting of a more complicated and general MIMO system model does not always hold for single-input systems. From a practical viewpoint, MIMO technology is becoming an integral part and a key feature of the next generation wireless communication systems. Thus, these results provide analytically justified design guidelines to maximize data rates subject to feedback constraints in densely populated urban areas with 4G communication systems.

In the second example above, the rate loss due to use of the homogenous feedback policy seems to be very minor around 0.01 [nats per channel use], and therefore it can be thought to be negligible for all practical purposes. This motivates us to examine the rate difference between homogenous and optimal threshold feedback policies for a broad spectrum of the SNR parameter to verify or falsify the validity of this conception. To this end, we investigate the optimality gap arising from the use of homogenous threshold

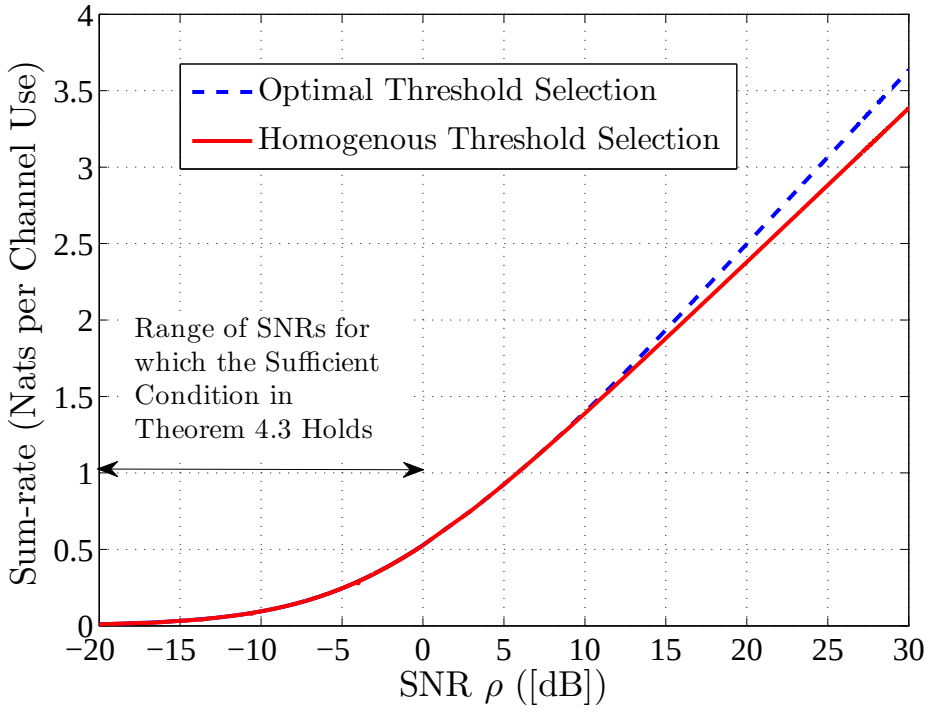


Figure 4.5: The optimality gap arising from the use of homogenous threshold feedback policies for different values of ρ when $M = 1$.

feedback policies as opposed to choosing thresholds optimally to maximize the sum-rate in Rayleigh fading environments in Fig. 4.5. We set n to 2, λ to 0.5 and M to 1 in this numerical example. Note that homogenous threshold feedback policies are always optimal when $M > 1$. Hence, there is no optimality gap to investigate in this case. For other values of λ and n, qualitatively similar observations continue to hold. Since we find optimal threshold levels through an exhaustive search, setting n to 2 limits our search space.

For small values of ρ up to 0 [dB], the homogenous threshold feedback policy with threshold levels set as $\boldsymbol{\tau}_{\text{homo}} = (\rho \log(4), \rho \log(4))^{\top}$ is optimum as predicted by Theorem 4.3. It continues to be optimum for a little while up to around 5.7 [dB] SNR values, and after which it becomes strictly suboptimal to use the homogenous threshold feedback policy in terms of the achieved downlink sum-rate. Furthermore, as channel conditions become better, *i.e.,* large values of ρ, the optimality gap becomes larger. Practically, this observation indicates that the use of homogenous threshold feedback policies may lead

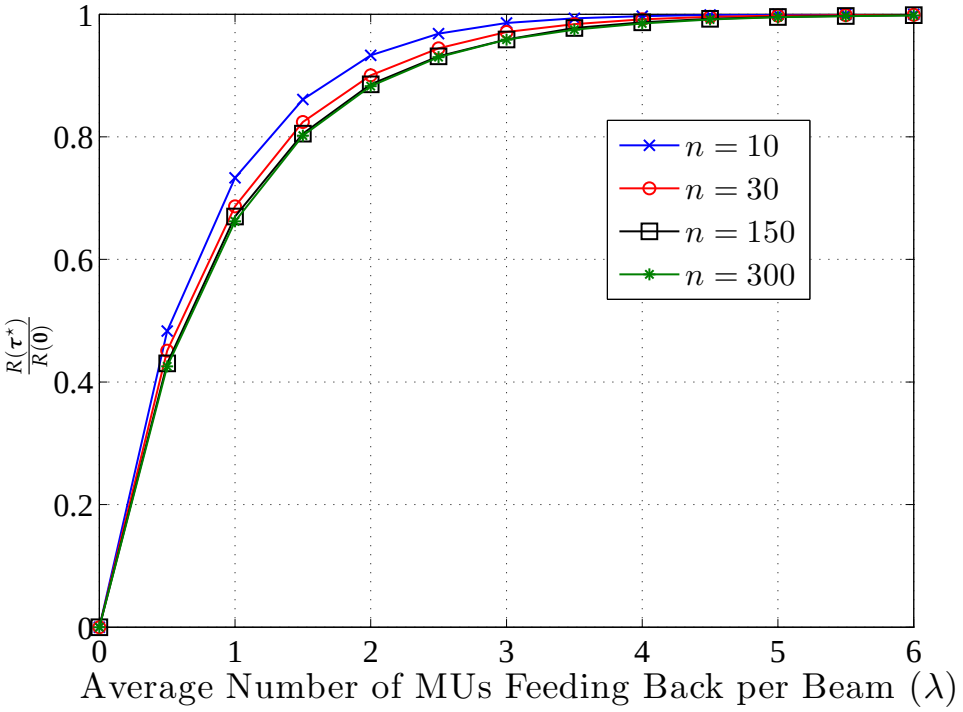


Figure 4.6: The ratio between the rates achieved with and without thresholding as a function of the average number of users feeding back per beam for different numbers of MUs.

to excessive rate loss in the high SNR regime for single beam systems when compared to the rate achieved by the optimal feedback policy.

Another important issue to investigate is the amount of feedback reduction that can be achieved by setting thresholds optimally. In Fig. 4.6, we plot the ratio $\frac{R(\boldsymbol{\tau}^\star)}{R(\mathbf{0})}$ between the rates achieved with and without thresholding as a function of λ for different numbers of MUs. In this figure, we set M to 1, and ρ to 1. Again, similar observations continue to hold for other values of M and ρ. Since $\rho = 1$, the homogenous threshold feedback policy with thresholds set as

$$\boldsymbol{\tau}^\star = \left(\rho \log\left(\frac{n}{\lambda}\right), \ldots, \rho \log\left(\frac{n}{\lambda}\right)\right)$$

is optimum, *i.e.*, see Lemma 4.8 and Theorem 4.3. After inspecting the figure, we see that there is almost no rate loss if the average number of MUs feeding back per beam is around five. We call this critical feedback level λ_c, which is an important design parameter to be

inputed to the higher MAC layer. It is interesting to see that the same design parameter applies to all $\frac{R(\boldsymbol{\tau}^\star)}{R(\mathbf{0})}$ curves that shift to the right only slightly and converge pointwise to a limiting curve as the number of MUs in the system increases.

The reason behind this phenomenon can be explained as follows. No feedback outage event occurs and beams are always assigned to the best MUs at each fading state when thresholds are set to zero. On the other hand, the feedback outage event probability is strictly positive when thresholds are optimally set to meet the feedback constraint λ. However, the tails of the distribution of the random number of MUs requesting each beam decays to zero exponentially fast, and therefore we are *almost* always guaranteed to have at least one MU demanding each beam whenever λ is above the critical feedback level λ_c. As a result, the feedback outage event probability becomes negligible, and the beams are still assigned to the best MUs with very high probability whenever $\lambda \geq \lambda_c$. Moreover, the distribution of the random number of MUs feeding back converges to a limiting distribution linearly with the total number of MUs in the system, which results in the observed pointwise convergence behavior in Fig. 4.6. Further details about the limiting $\frac{R(\boldsymbol{\tau}^\star)}{R(\mathbf{0})}$ curve (as $n \to \infty$) can be found in Chapter 5, where its exact characterization is obtained and interpreted as the *feedback-capacity tradeoff curve*.

Two possible interpretations about λ_c are as follows. Since the base station communicates only with the best MU on each beam, an ideal feedback policy in terms of the optimal usage of uplink communication resources is the one that only allows the best MU to feed back at each channel fading state. However, such a policy requires centralized operation, or coordination among MUs. Thus, when compared with the ideal feedback policy, λ_c can be interpreted as the price that we have to pay to achieve *almost* the same performance with the ideal feedback policy due to decentralized operation. Secondly, when compared with the all-feedback policy, it represents the amount of feedback reduction that can be achieved without any noticeable performance degradation. For example, as opposed to allowing all MUs to feed back, we can reduce the total feedback load 30 times and 60 times by setting thresholds optimally when $n = 150$ and $n = 300$, respectively, without any evident performance loss.

4.6.2 Rician and Nakagami Fading Channels

In this part, we extend our analysis above to other channel models by briefly studying optimality and sub-optimality regions for homogenous threshold feedback policies for Nakagami and Rician fading channel models. We set M to 1 for simplicity. Otherwise, calculations for the $M > 1$ case easily gets very complicated for these channel models, which hinders the intuitive understanding of the results below. In particular, the derivation of the SINR distribution in the general case becomes very complex.

We start our discussion with Nakagami fading channels. In this case, $h_{k,i}, k = 1, \ldots, N_t$ and $i = 1, \ldots, n$, are i.i.d. with the common distribution Nakagami (μ, ω), where μ and ω are shape and spread parameters, respectively. Hence, channel power gains are Gamma distributed with distribution Gamma $\left(\mu, \frac{\omega}{\mu}\right)$, where μ and $\frac{\omega}{\mu}$ are shape and scale parameters of the associated Gamma distribution, respectively. We first note that ω is equal to the average channel power gain, and therefore it is set to 1 to be consistent with the Rayleigh fading channel model above. Secondly, if X is a random variable with distribution Gamma $\left(\mu, \frac{1}{\mu}\right)$, then aX is distributed according to Gamma $\left(\mu, \frac{a}{\mu}\right)$, where a is a positive real number. Therefore, under this channel model, the SINR[5] distribution is equal to Gamma $\left(\mu, \frac{\rho}{\mu}\right)$.

In Fig. 4.7, we illustrate the regions on which homogenous threshold feedback policies are optimal and suboptimal for the Nakagami fading channel model. We set n to 2 and λ to 0.5 in this figure. The same observations continue hold for other parameter selections. The blue region is computed numerically by using the sufficient condition for the Schur-concavity of the sum-rate in Theorem 4.1, whereas the red region is obtained by evaluating the sufficient condition in Theorem 4.2 numerically. As mentioned earlier, the sufficient condition in Theorem 4.1 is stronger than the one in Theorem 4.2, which is why the red region is contained within the blue region in Fig. 4.7. Note that the Nakagami fading model reduces to the Rayleigh fading model, and the red region only covers SNR values less than one when $\mu = 1$, which is in accordance with our discussion and Theorem 4.3 above. Surprisingly, our numerical investigation shows that homogenous

[5]No inter-beam interference exists in the $M = 1$ case. Hence, the random SINR is the same quantity with the random SNR. We continue to use the term SINR for this case to avoid any confusion with the average SNR ρ.

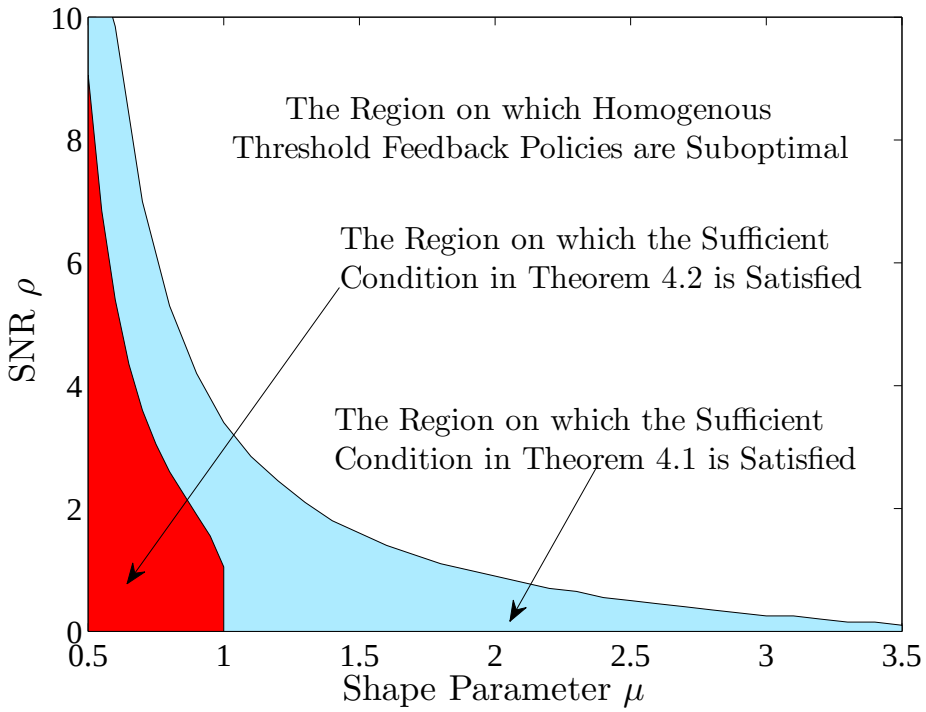


Figure 4.7: The regions on which homogenous threshold feedback policies are optimal and suboptimal for the Nakagami fading channel model.

threshold feedback policies are suboptimal outside the blue region in Fig. 4.7. Therefore, we conjecture that the condition provided in Theorem 4.1 is also necessary for the optimality of homogenous threshold feedback policies.

Secondly, we consider the Rician fading channel model in which the channel amplitude gains are Rician distributed with distribution Rician (K, P), where P is the total power gain and K (*a.k.a.*, K factor) is the ratio between the power in the direct path and the power in the scattered paths. We set P to 1 to be consistent with the Rayleigh and Nakagami fading channel models studied above. If X is a random variable with distribution Rician (K, P), then $\left(\frac{X}{\sigma}\right)^2$ has a non-central Chi-square distribution with two degrees of freedom, and the non-centrality parameter is given by $2K$ if the scaling coefficient σ is chosen to be $\sigma = \sqrt{\frac{P}{2(1+K)}}$. We obtain the SINR distribution by scaling this non-central Chi-square distribution with $\rho\sigma^2$.

Fig. 4.8 illustrates the regions on which homogenous threshold feedback policies are

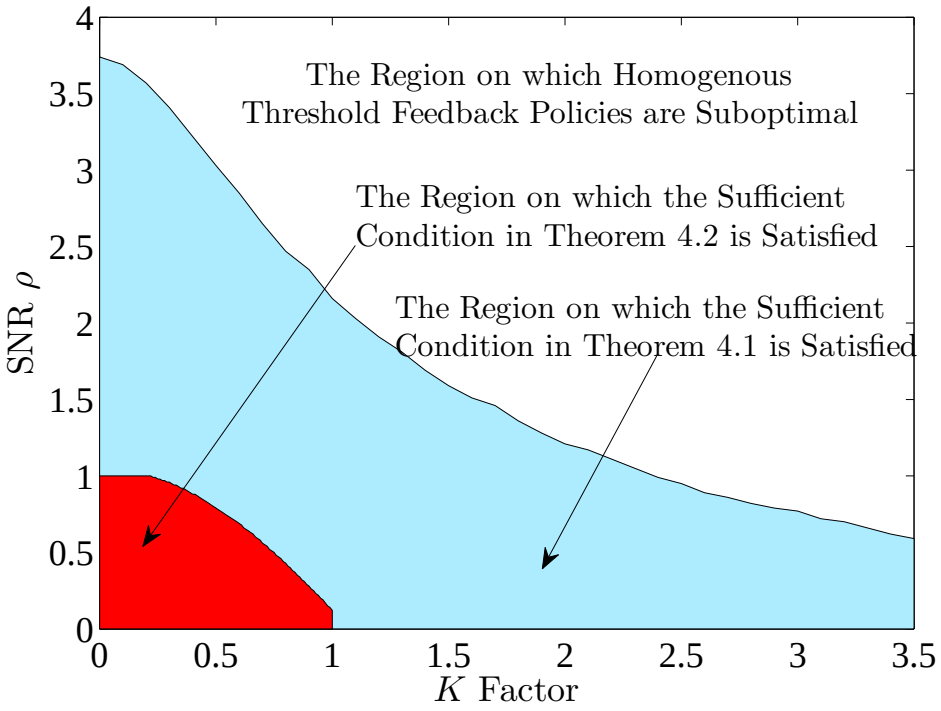


Figure 4.8: The regions on which homogenous threshold feedback policies are optimal and suboptimal for the Rician fading channel model.

optimal and suboptimal for the Rician fading channel model. We set n to 2 and λ to 0.5 in this figure. Since the similar explanations above continue to hold for the Rician fading channel model as well, we do not repeat them here again.

4.6.3 Why Does Sub-optimality Arise?

In this part, we provide an intuitive explanation for why homogenous threshold feedback policies sometimes become suboptimal to use even when MUs experience statistically the same channel conditions. Our discussion will focus on the single beam case first.

Let β be the feedback outage event probability, $R\left(\boldsymbol{\tau}_{\text{homo}}\right)$ be the sum-rate achieved by the homogenous threshold feedback policy satisfying feedback constraints with equality, and $R\left(\boldsymbol{\tau}^{\star}\right)$ be the sum-rate achieved by setting thresholds optimally. For simplicity, we let $n = 2$, but similar explanations continue to hold for any n. The sum-rate in this case

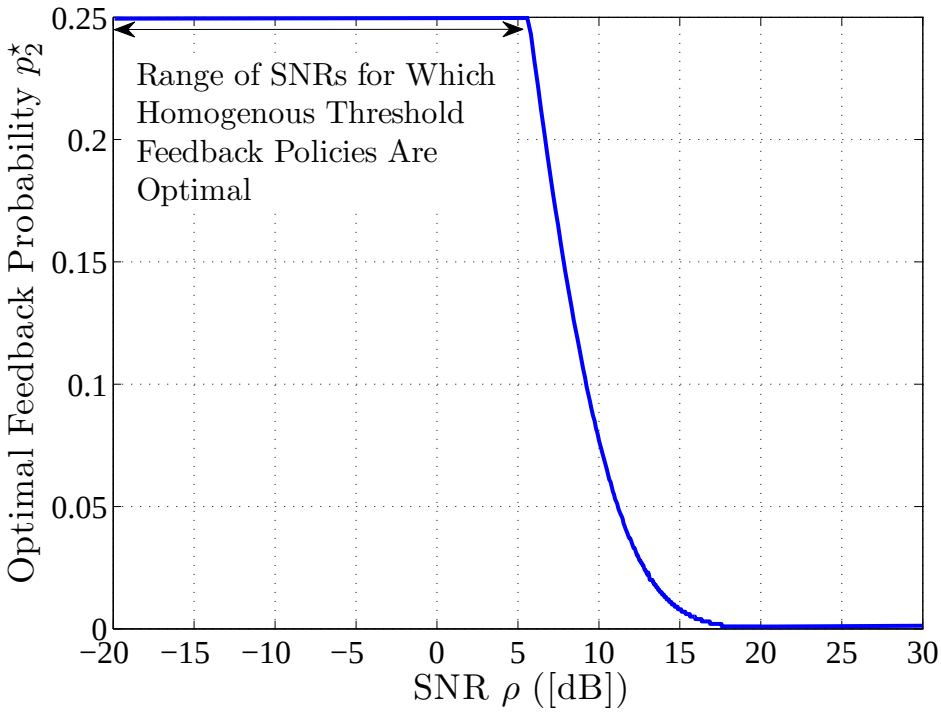


Figure 4.9: Optimal feedback probability $p_2^\star$ of the second MU as a function of ρ. ($\lambda = 0.5$)

can be written as

$$R(\tau_1, \tau_2) = (1-\beta)\,\mathsf{E}\left[\log\left(1+\max_{i=1,2}\gamma_i 1_{\{\gamma_i \geq \tau_i\}}\right)\Big|\ \text{No Outage}\right].$$

Two key underlying factors affect this rate expression. The first one is the *power gain* that can be achieved by means of multiuser diversity. This is represented by the maximization operation inside the logarithm function above. The more MUs feed back, the more likely the output of this maximization operation to be higher. Indeed, the exact asymptotic statistics of the resulting power gain (under various channel models) can be obtained by resorting to an order statistics analysis [19]. The second factor is the *degrees-of-freedom gain* represented by the $1-\beta$ term. The smaller the feedback outage event probability, the higher the degrees-of-freedom gain that we achieve. The choice of thresholds affects both gains, and the interplay between them determines how we set thresholds to maximize the downlink sum-rate.

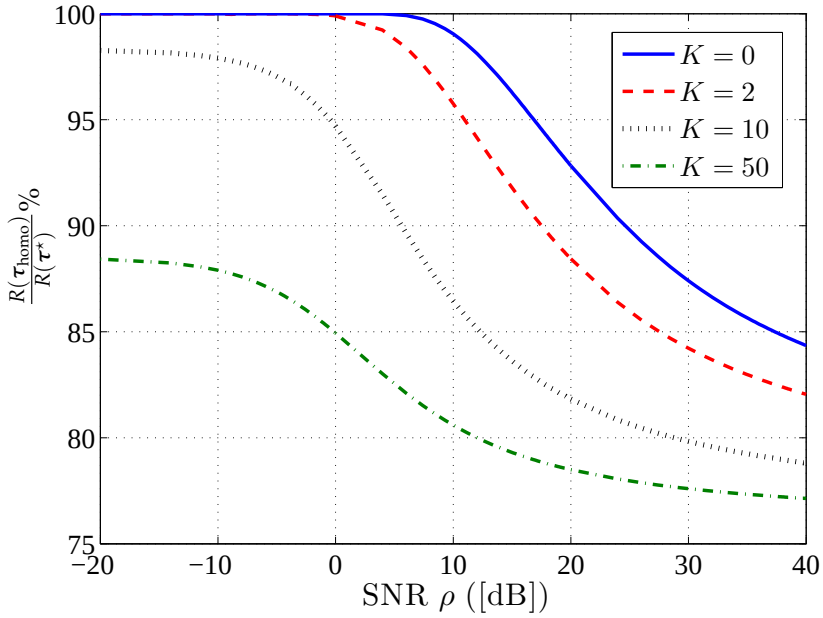
In Fig. 4.9, we focus on the Rayleigh fading channel model to provide further details

about the interplay between power and degrees-of-freedom gains. In this figure, we set λ to 0.5, and plot the optimal feedback probability $p_2^\star$ of the second MU as a function of ρ. In the low SNR regime, $p_2^\star$ is equal to 0.25, which implies the optimality of the homogenous threshold feedback policy equalizing the feedback probabilities of both MUs. However, as ρ increases, we start to prefer one MU over the other one to maximize the sum-rate. In this case, for example, we prefer the first MU over the second one by decreasing the feedback probability of the second MU to zero, and increasing the feedback probability of the first MU to 0.5 in the high SNR regime.

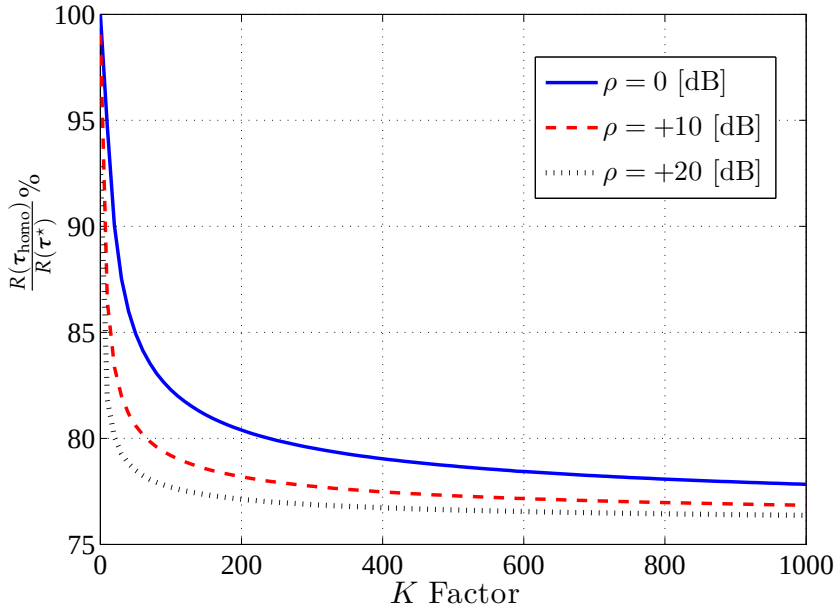
The main reason behind this behavior is as follows. When the SNR is low, the sum-rate increases *almost* linearly with the power gain. As a result, we tend to choose thresholds equally to maximize the power gain, and thereby to maximize the sum-rate, in the low SNR regime although such a threshold assignment reduces the degrees-of-freedom gain. In the high SNR regime, on the other hand, the power gain can only provide a logarithmic increase in the sum-rate, *i.e.,* the law of diminishing returns. Hence, the power gain earned by setting thresholds equally becomes negligible when compared to the loss in the degrees-of-freedom gain, and we tend to choose thresholds heterogeneously to maximize the degrees-of-freedom gain, and thereby to maximize the sum-rate, in the high SNR regime. A similar behavior continues to hold for other channel models, which is what we investigate next.

In Figs. 4.10(a) and 4.10(b), we plot the ratio $\frac{R(\boldsymbol{\tau}_{\text{homo}})}{R(\boldsymbol{\tau}^\star)}$ as a function of ρ and K, respectively, for the Rician fading channel model. We set λ to 1 in both figures. The SNR has the same effect on how we set thresholds optimally in the Rician case as well. When small, we prefer the power gain over the degrees-of-freedom gain, and set thresholds equally to maximize the sum-rate, which is why $\frac{R(\boldsymbol{\tau}_{\text{homo}})}{R(\boldsymbol{\tau}^\star)}$ ratio is around one for small values of ρ, and for $K = 0$ and 2. When high, we prefer the degrees-of-freedom gain over the power gain, and set thresholds unequally to maximize the sum-rate, which is why $\frac{R(\boldsymbol{\tau}_{\text{homo}})}{R(\boldsymbol{\tau}^\star)}$ ratio converges to 0.75 for high values of ρ.

The exact behavior of $\frac{R(\boldsymbol{\tau}_{\text{homo}})}{R(\boldsymbol{\tau}^\star)}$ strongly depends on K, too. Roughly speaking, K determines the dynamic range of the SINR distribution, and the power gain due to multiuser diversity becomes more prominent when the dynamic range of the distribution is large



(a) As a function of ρ.



(b) As a function of K.

Figure 4.10: The change of the ratio between the sum-rates achieved by homogenous and optimal threshold feedback policies as a function of ρ and K for a Rician fading model. ($\lambda = 1$)

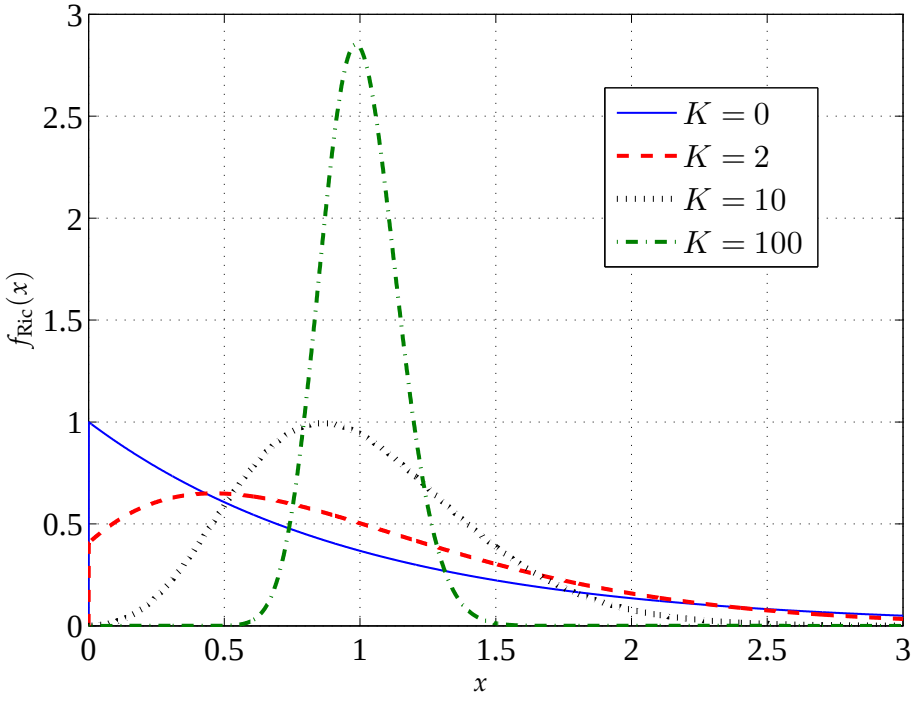
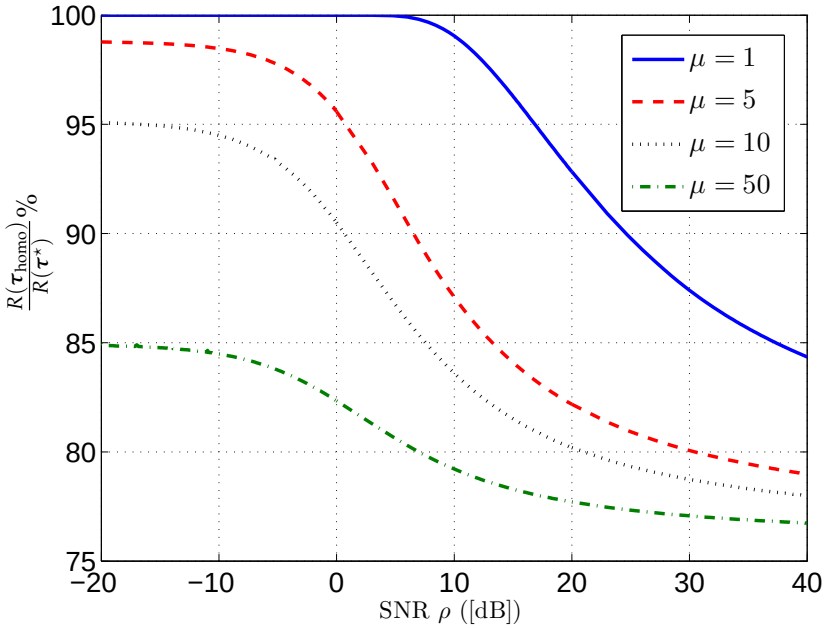


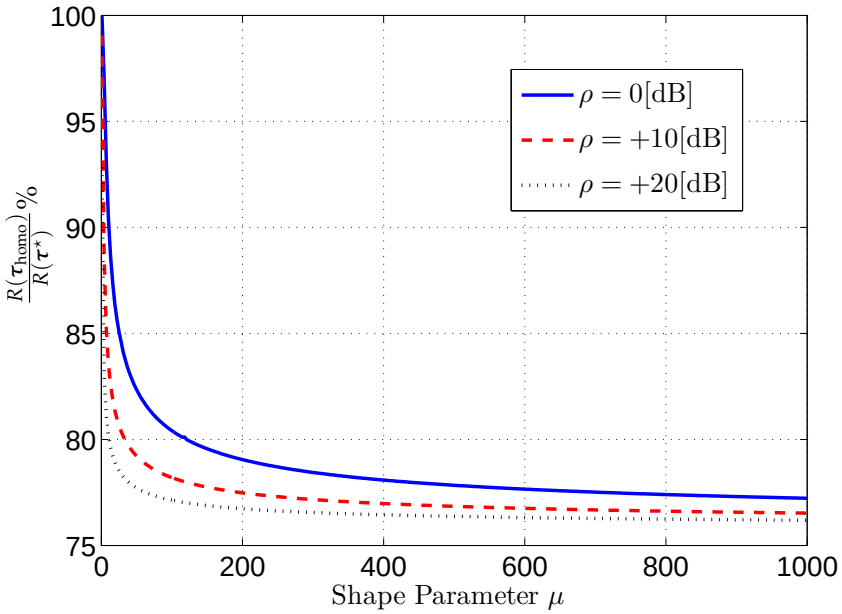
Figure 4.11: Dynamic range of the SINR distribution for the Rician fading channel model for different values of K. ($\rho = 1$)

[91]. However, as K increases, the power in the direct path increases, which, in turn, nullifies the scattering effects and reduces the dynamic range of the SINR distribution, *e.g.*, see Fig. 4.11 for an illustration. Therefore, regardless of how small the SNR is, it may still become suboptimal to use homogenous threshold feedback policies when K is large, as illustrated by the curves corresponding to $K = 10$ and 50 in Fig. 4.10(a). Furthermore, as K increases, the channel becomes more deterministic, and we experience almost no power gain due to multiuser diversity in the limit. As a result, $\frac{R(\boldsymbol{\tau}_{\text{homo}})}{R(\boldsymbol{\tau}^\star)}$ still converges to 0.75 as K grows large, which is illustrated by Fig. 4.10(b).

Finally, we note that the limiting value of $\frac{R(\boldsymbol{\tau}_{\text{homo}})}{R(\boldsymbol{\tau}^\star)}$ (in the high SNR, or high K regime) depends on the feedback constraint λ. If $\lambda \leq 1$, the optimum feedback probability selection converges to $p_1 = \lambda$ and $p_2 = 0$ (or, vice versa) when ρ or K grows large. Hence, $\frac{R(\boldsymbol{\tau}_{\text{homo}})}{R(\boldsymbol{\tau}^\star)}$ converges to $1 - \frac{\lambda}{4}$ for $\lambda \leq 1$, which is inline with the 0.75 limit to which the curves in both Figs. 4.10(a) and 4.10(b) converge. If $\lambda > 1$, the optimum feedback probability selection converges to $p_1 = 1$ and $p_2 = \lambda - 1$ (or, vice versa) when ρ or K grows



(a) As a function of ρ.



(b) As a function of μ.

Figure 4.12: The change of the ratio between the sum-rates achieved by homogenous and optimal threshold feedback policies as a function of ρ and μ for a Nakagami fading model. ($\lambda = 1$)

large. Hence, $\frac{R(\boldsymbol{\tau}_{\text{homo}})}{R(\boldsymbol{\tau}^\star)}$ converges to $\lambda - \frac{\lambda^2}{4}$ for $\lambda > 1$. Let $C^\star(\lambda)$, $0 \leq \lambda \leq 2$, be the limiting value that $\frac{R(\boldsymbol{\tau}_{\text{homo}})}{R(\boldsymbol{\tau}^\star)}$ converges as ρ or K grows large. It is not hard to see that the minimum value of $C^\star(\lambda)$ is 0.75, which is achieved when $\lambda = 1$.

A similar analysis can be done for a Nakagami channel model as well. In Figs. 4.12(a) and 4.12(b), we plot the ratio $\frac{R(\boldsymbol{\tau}_{\text{homo}})}{R(\boldsymbol{\tau}^\star)}$ as a function of ρ and μ, respectively. λ is set at 1 in both figures. We omit the explanation to avoid repetition. Using the analysis on the Rician and Nakagami models, we can obtain the maximum optimality loss arising from the use of homogenous threshold feedback policies for a two-user single beam system, which we state through the following corollary.

Corollary 4.1. *The maximum optimality loss arising from the use of homogenous threshold feedback policies for a two-user single beam system is* 25%.

Up to now, we have only focused on the single beam case to explain why homogenous threshold feedback policies may sometimes become suboptimal to use. Based on the arguments above, we provide further insights as to why homogenous threshold feedback policies are always optimal to use when $M > 1$ for the Rayleigh fading channel model. We first note that, in contrast to the single beam case, Theorem 4.4 indicates a potential phase transition phenomenon in the behavior of the sum-rate in which homogenous threshold feedback policies suddenly become always optimal to use when we go from the single beam case to the multiple beams case. The main reason behind this phenomenon is the inter-beam interference when multiple beams are used to communicate with multiple MUs simultaneously. Such a multiuser operation makes the network interference limited, rather than being noise limited, when compared to the single beam case. More specifically, an increase in SNR implies a corresponding increase in the inter-beam interference experienced by other beams, and the system ends up operating always in the low SNR regime effectively when $M > 1$. Therefore, the low SNR Rayleigh fading behavior kicks in, and homogenous threshold feedback policies always become optimal to use. On the other hand, received signal powers improve linearly with SNR in the single beam case, which makes homogenous threshold feedback policies suboptimal to use in the high SNR regime.

Although this intuition works for the Rayleigh fading channel model, it is too op-

timistic to ask for the optimality of homogenous threshold feedback policies for other channel models as well when $M > 1$. As our discussion above makes it apparent, the power gain due to multiuser diversity strongly depends on the parameters of the fading process determining the dynamic range of the resulting SINR distribution. There is no power gain to benefit from multiuser diversity by giving all MUs equal chances of channel access if the SINR distribution becomes increasingly more deterministic. In these instances, it is expected that a heterogenous threshold assignment will maximize the sum-rate even if the network is interference limited due to multi-beam operation. It is a potential future research interest to investigate the conditions on the parameters of the fading process to guarantee the optimality of homogenous threshold feedback policies for channel models other than the Rayleigh fading model such as Rician and Nakagami fading channels.

4.7 Conclusions

Having established the optimality of threshold feedback policies in the previous chapter, we now faced an optimal threshold selection problem to maximize the sum-rate. This is a non-convex optimization problem over finite dimensional Euclidean spaces. We solved this problem by identifying an underlying Schur-concave structure in the sum-rate when it is viewed as a function of feedback probabilities. Specifically, we have obtained sufficient conditions ensuring the Schur-concavity of the sum-rate, and therefore the rate optimality of homogenous threshold feedback policies in which all MUs use the same threshold for their feedback decisions. These sufficient conditions have been provided for general fading channel models as well.

We have performed an extensive numerical and simulation study to illustrate the applications of our results to familiar fading channel models such as Rayleigh, Nakagami and Rician fading channels. With some surprise, we have shown that homogenous threshold feedback policies are not always optimal to use for general fading channels, even when all MUs experience statistically the same channel conditions. In the particular case of Rayleigh fading channels, on the other hand, homogenous threshold feedback

policies have been proven to be rate-wise optimal if multiple beams are used for the downlink communication. We have also studied the optimality and sub-optimality regions for the homogenous threshold feedback policies in the Rician and Nakagami case. The detailed insights regarding when and why homogenous threshold feedback policies are rate-wise optimal or suboptimal have been provided, in conjunction with various other design and engineering perspectives.

4.8 Appendix

4.8.1 Proof of Lemma 4.2

Assume $\tau_2 \geq \tau_1$ (*i.e.*, $\tau_1 = \tau_{\pi(1)}$ and $\tau_2 = \tau_{\pi(2)}$) for notational simplicity. Then, for a two-user system, the rate on beam 1 as a function of the thresholds is given as

$$\begin{aligned} R^1(\tau_1,\tau_2) &= F(\tau_1)\int_{\tau_2}^{\infty}\log(1+x)dF(x) + F(\tau_2)\int_{\tau_1}^{\infty}\log(1+x)dF(x) \\ &+ \mathsf{E}\left[\log\left(1+\max\{\gamma_{1,1},\gamma_{1,2}\}\right)1_{\{\gamma_{1,1}\geq\tau_1,\gamma_{1,2}\geq\tau_2\}}\right] \\ &= F(\tau_1)\int_{\tau_2}^{\infty}\log(1+x)dF(x) + F(\tau_2)\int_{\tau_1}^{\infty}\log(1+x)dF(x) \\ &+ (1-F(\tau_1))(1-F(\tau_2))\,\mathsf{E}\left[\log\left(1+\max\{\gamma_{1,1},\gamma_{1,2}\}\right)|\gamma_{1,1}\geq\tau_1,\gamma_{1,2}\geq\tau_2\right]. \end{aligned}$$

Let

$$H(x) = \mathsf{Pr}\left\{\max\{\gamma_{1,1},\gamma_{1,2}\}\leq x|\gamma_{1,1}\geq\tau_1,\gamma_{1,2}\geq\tau_2\right\},$$

i.e., $H(x)$ is the CDF of $\max\{\gamma_{1,1},\gamma_{1,2}\}$ given $\gamma_{1,1}\geq\tau_1$ and $\gamma_{1,2}\geq\tau_2$. Then,

$$H(x) = \begin{cases} \frac{F(x)-F(\tau_1)}{1-F(\tau_1)}\cdot\frac{F(x)-F(\tau_2)}{1-F(\tau_2)} & \text{if } x\geq\max\{\tau_1,\tau_2\} \\ 0 & \text{if } x<\max\{\tau_1,\tau_2\} \end{cases}. \tag{4.17}$$

We can write $R^1(\tau_1,\tau_2)$ as

$$R^1(\tau_1,\tau_2) = F(\tau_1)\int_{\tau_2}^{\infty}\log(1+x)dF(x) + F(\tau_2)\int_{\tau_1}^{\infty}\log(1+x)dF(x)$$

$$+ \quad (1-F(\tau_1))(1-F(\tau_2)) \int_{\max\{\tau_1,\tau_2\}}^{\infty} \log(1+x) dH(x), \tag{4.18}$$

and substituting (4.17) in (4.18) leads to

$$R^1(\tau_1,\tau_2) = \int_{\tau_2}^{\infty} \log(1+x) dF^2(x) + F(\tau_2) \int_{\tau_1}^{\tau_2} \log(1+x) dF(x), \tag{4.19}$$

for $\tau_2 \geq \tau_1$. For $\tau_1 \geq \tau_2$, we just switch the places of τ_1 and τ_2 in (4.19). Hence, the proof is complete.

4.8.2 Rate for Different Values of $\bar{\gamma}^{\star}_{\mathcal{N}'}$

Proof of Lemma 4.3

Let $\bar{\xi}^{\star}_{i+1,i} = \max\left\{\bar{\gamma}_{1,\pi(i+1)}, \bar{\gamma}_{1,\pi(i)}\right\}$. From (4.11),

$$R^1\left(\tau_{\pi(i+1)}, \tau_{\pi(i)} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right) = \log\left(1+\bar{\gamma}^{\star}_{\mathcal{N}'}\right) \mathsf{Pr}\left\{\bar{\xi}^{\star}_{i+1,i} \leq \bar{\gamma}^{\star}_{\mathcal{N}'} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right\} + \mathsf{E}\left[\log\left(1+\bar{\xi}^{\star}_{i+1,i}\right) 1_{\left\{\bar{\xi}^{\star}_{i+1,i} > \bar{\gamma}^{\star}_{\mathcal{N}'}\right\}} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right]. \tag{4.20}$$

Let $\mathcal{A} = \left\{\bar{\xi}^{\star}_{i+1,i} \leq \bar{\gamma}^{\star}_{\mathcal{N}'}\right\}$ and $\mathcal{B} = \left\{\xi^{\star}_{i+1,i} \leq \bar{\gamma}^{\star}_{\mathcal{N}'}\right\}$. Since $\bar{\gamma}^{\star}_{\mathcal{N}'}$ is larger than $\tau_{\pi(i+1)}$, it follows that $\mathcal{A} = \mathcal{B}$. Thus, we can write

$$\mathsf{Pr}\left\{\bar{\xi}^{\star}_{i+1,i} \leq \bar{\gamma}^{\star}_{\mathcal{N}'} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right\} = \mathsf{Pr}\left\{\xi^{\star}_{i+1,i} \leq \bar{\gamma}^{\star}_{\mathcal{N}'} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right\}$$

for the first term on the righthand side of (4.20). For the second term, we have $\bar{\xi}^{\star}_{i+1,i} = \xi^{\star}_{i+1,i}$ since $\bar{\xi}^{\star}_{i+1,i} > \bar{\gamma}^{\star}_{\mathcal{N}'} > \tau_{\pi(i+1)}$, which concludes the proof.

Proof of Lemma 4.4

When $\bar{\gamma}^{\star}_{\mathcal{N}'} < \tau_{\pi(i)}$, (4.20) simplifies to

$$R^1\left(\tau_{\pi(i+1)}, \tau_{\pi(i)} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right) = \mathsf{E}\left[\log\left(1+\xi^{\star}_{i+1,i}\right) 1_{\left\{\gamma_{1,\pi(i)} > \tau_{\pi(i)}, \gamma_{1,\pi(i+1)} > \tau_{\pi(i+1)}\right\}}\right] + \mathsf{E}\left[\log\left(1+\gamma_{1,\pi(i+1)}\right) 1_{\left\{\gamma_{1,\pi(i+1)} > \tau_{\pi(i+1)}\right\}}\right] F\left(\tau_{\pi(i)}\right)$$

$$+ \mathsf{E}\left[\log\left(1+\gamma_{1,\pi(i)}\right) 1_{\left\{\gamma_{1,\pi(i)}>\tau_{\pi(i)}\right\}}\right] F\left(\tau_{\pi(i+1)}\right) + \log\left(1+\bar{\gamma}^{\star}_{\mathcal{N}'}\right) F\left(\tau_{\pi(i)}\right) F\left(\tau_{\pi(i+1)}\right).$$

The first three terms on the righthand side is identical to the rate expression for the two-user system in Lemma 4.2. Substituting the result for the two-user case completes the proof.

Proof of Lemma 4.5

For $\tau_{\pi(i)} \leq \bar{\gamma}^{\star}_{\mathcal{N}'} \leq \tau_{\pi(i+1)}$, (4.20) simplifies to

$$\begin{aligned} R\left(\tau_{\pi(i+1)}, \tau_{\pi(i)} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right) &= \log\left(1+\bar{\gamma}^{\star}_{\mathcal{N}'}\right) F\left(\tau_{\pi(i+1)}\right) F\left(\bar{\gamma}^{\star}_{\mathcal{N}'}\right) \\ &+ \mathsf{E}\left[\log\left(1+\gamma_{1,\pi(i)}\right) 1_{\left\{\gamma_{1,\pi(i)}>\bar{\gamma}^{\star}_{\mathcal{N}'}\right\}} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right] F\left(\tau_{\pi(i+1)}\right) \\ &+ \mathsf{E}\left[\log\left(1+\gamma_{1,\pi(i+1)}\right) 1_{\left\{\gamma_{1,\pi(i+1)}>\tau_{\pi(i+1)}\right\}}\right] F\left(\bar{\gamma}^{\star}_{\mathcal{N}'}\right) \\ &+ \mathsf{E}\left[\log\left(1+\xi^{\star}_{i+1,i}\right) 1_{\left\{\gamma_{1,\pi(i)}>\bar{\gamma}^{\star}_{\mathcal{N}'}, \gamma_{1,\pi(i+1)}>\tau_{\pi(i+1)}\right\}} | \bar{\gamma}^{\star}_{\mathcal{N}'}\right] \end{aligned}$$

The last three terms on the righthand side can be further simplified as in Lemma 4.2 for the two-user system, which completes the proof.

4.8.3 Proof of Theorem 4.1

It is enough to show that $R^1\left(q|\bar{\gamma}^{\star}_{\mathcal{N}'}\right)$ is a non-decreasing function of $q \in \mathcal{P}_{i+1}$ for all $i = 1, \ldots, n-1$ and $\bar{\gamma}^{\star}_{\mathcal{N}'} \geq 0$ based on Lemma 4.1. To this end, we can write $R_1\left(q|\bar{\gamma}^{\star}_{\mathcal{N}'}\right)$ explicitly as

$$\begin{aligned} R_1\left(q|\bar{\gamma}^{\star}_{\mathcal{N}'}\right) = \int_{F^{-1}(1-q)}^{\infty} \log(1+x) dF^2(x) + (1-q) \int_{F^{-1}(1+q-\lambda_i)}^{F^{-1}(1-q)} \log(1+x) dF(x) \\ + \log\left(1+\bar{\gamma}^{\star}_{\mathcal{N}'}\right)(1-q)(1+q-\lambda_i). \end{aligned}$$

Using Lemma 4.7, we get

$$\frac{dR_1\left(q|\bar{\gamma}^{\star}_{\mathcal{N}'}\right)}{dq} = \log\left(1+\bar{\gamma}^{\star}_{\mathcal{N}'}\right)(\lambda_i - 2q)$$

$$+\int_{F^{-1}(1+q-\lambda_i)}^{F^{-1}(1-q)} \frac{F(x)}{1+x}dx-(\lambda_i-2q)\log\left(1+F^{-1}(1+q-\lambda_i)\right). \quad (4.21)$$

Similarly, we can write $R_2\left(q|\bar{\gamma}^{\star}_{\mathcal{N}'}\right)$ explicitly as

$$R_2\left(q|\bar{\gamma}^{\star}_{\mathcal{N}'}\right)=\int_{F^{-1}(1-q)}^{\infty}\log(1+x)dF^2(x)+(1-q)\int_{\bar{\gamma}^{\star}_{\mathcal{N}'}}^{F^{-1}(1-q)}\log(1+x)dF(x)$$
$$+\log\left(1+\bar{\gamma}^{\star}_{\mathcal{N}'}\right)(1-q)\,F\left(\bar{\gamma}^{\star}_{\mathcal{N}'}\right).$$

Differentiation and integration-by-parts give us

$$\frac{dR_2\left(q|\bar{\gamma}^{\star}_{\mathcal{N}'}\right)}{dq}=\int_{\bar{\gamma}^{\star}_{\mathcal{N}'}}^{F^{-1}(1-q)}\frac{F(x)}{1+x}dx\geq 0.$$

Thus, $R^1\left(q|\bar{\gamma}^{\star}_{\mathcal{N}'}\right)$ is a non-decreasing function of $q\in\mathcal{P}_{i+1}$ for all $i=1,\ldots,n-1$ and $\bar{\gamma}^{\star}_{\mathcal{N}'}\geq 0$ if (4.6) is correct.

4.8.4 Proof of Theorem 4.2

Let

$$U\left(q,\lambda\right)=\int_{F^{-1}(1+q-\lambda)}^{F^{-1}(1-q)}\frac{F(x)}{1+x}dx-(\lambda-2q)\log\left(1+F^{-1}\left(1+q-\lambda\right)\right).$$

Then, it is enough to show that $U\left(q,\lambda\right)\geq 0$ for all $\lambda\in[0,2]$ and $\max\left\{0,\lambda-1\right\}\leq q\leq\frac{\lambda}{2}$ by Theorem 4.1. To this end, it is enough to show $\frac{\partial U(q,\lambda)}{\partial q}\leq 0$ for all $\lambda\in[0,2]$ and $\max\left\{0,\lambda-1\right\}\leq q\leq\frac{\lambda}{2}$ since $U\left(\frac{\lambda}{2},\lambda\right)=0$.

The following lemma simplifies the proof considerably.

Lemma 4.9. *If f is bounded at zero and f' satisfies $f'\left(F^{-1}(x)\right)\leq-\frac{f\left(F^{-1}(x)\right)}{1+F^{-1}(x)}$ for all $x\in[0,1]$, then $G\leq 0$ on $[0,1]$, where*

$$G(x)=\log\left(1+F^{-1}(x)\right)\left(1+F^{-1}(x)\right)f\left(F^{-1}(x)\right)-x$$

for $x\in[0,1]$.

Proof. By taking the first derivative of $G(x)$ with respect to x,

$$\begin{aligned}\frac{dG(x)}{dx} &= \log\left(1+F^{-1}(x)\right)\left(1+F^{-1}(x)\right)\frac{f'\left(F^{-1}(x)\right)}{f\left(F^{-1}(x)\right)}+\log\left(1+F^{-1}(x)\right)\\ &= \log\left(1+F^{-1}(x)\right)\left[1+\frac{\left(1+F^{-1}(x)\right)f'\left(F^{-1}(x)\right)}{f\left(F^{-1}(x)\right)}\right]<0.\end{aligned}$$

Hence, $G(x)$ is strictly decreasing for $x>0$, and achieves its maximum at $x=0$. We have $\lim_{x\to 0} G(x)=0$ since $f(x)$ is bounded at 0, which completes the proof. □

Now, consider the partial derivative of $U(q,\lambda)$ with respect to q, which is equal to

$$\begin{aligned}\frac{\partial U(q,\lambda)}{\partial q} &= \frac{1-q}{1+F^{-1}(1-q)}.\frac{-1}{f\left(F^{-1}(1-q)\right)}-\frac{1+q-\lambda}{1+F^{-1}(1+q-\lambda)}.\frac{1}{f\left(F^{-1}(1+q-\lambda)\right)}\\ &- \frac{\lambda-2q}{1+F^{-1}(1+q-\lambda)}.\frac{1}{f\left(F^{-1}(1+q-\lambda)\right)}+2\log\left(1+F^{-1}(1+q-\lambda)\right).\end{aligned}$$

Taking the common denominators gives us

$$\frac{\partial U(q,\lambda)}{\partial q}=K_1(q)\,g_1(q,\lambda)+K_2(q,\lambda)\,g_2(q,\lambda),$$

where

$$\begin{aligned}g_1(q,\lambda) &= \log\left(1+F^{-1}(1+q-\lambda)\right)\left(1+F^{-1}(1-q)\right)f\left(F^{-1}(1-q)\right)-(1-q),\\ g_2(q,\lambda) &= \log\left(1+F^{-1}(1+q-\lambda)\right)\left(1+F^{-1}(1+q-\lambda)\right)f\left(F^{-1}(1+q-\lambda)\right)-(1-q),\\ K_1(q) &= \frac{1}{\left(1+F^{-1}(1-q)\right)f\left(F^{-1}(1-q)\right)},\text{ and}\\ K_2(q,\lambda) &= \frac{1}{\left(1+F^{-1}(1+q-\lambda)\right)f\left(F^{-1}(1+q-\lambda)\right)}.\end{aligned}$$

Note that K_1 and K_2 are always positive. Thus, it is enough to show that g_1 and g_2 are non-positive on $\left[\max\{0,\lambda-1\},\frac{\lambda}{2}\right]$ for any fixed $\lambda\in[0,2]$. To this end, g_1 and g_2 on $\left[\max\{0,\lambda-1\},\frac{\lambda}{2}\right]$ can be upper bounded as

$$g_1(q,\lambda)\le g_1^u(q) = \log\left(1+F^{-1}(1-q)\right)\left(1+F^{-1}(1-q)\right)f\left(F^{-1}(1-q)\right)-(1-q)$$

and

$$g_2(q,\lambda) \le g_2^u(q,\lambda) = \log\left(1+F^{-1}(1+q-\lambda)\right) \times \left(1+F^{-1}(1+q-\lambda)\right) f\left(F^{-1}(1+q-\lambda)\right) - (1+q-\lambda).$$

Now, using Lemma 4.9, we can show that both g_1^u and g_2^u are non-positive functions on $\left[\max(0,\lambda-1), \frac{\lambda}{2}\right]$. This means $\frac{\partial U(q,\lambda)}{\partial q} \le 0$, which implies $U(q,\lambda) \ge 0$ for all $\lambda \in [0,2]$ and $\max\{0,\lambda-1\} \le q \le \frac{\lambda}{2}$.

4.8.5 Proof of Lemma 4.8

For $M = 1$, it is easy to get $F^{-1}(x) = -\rho\log(1-x)$. For $M > 1$, we need to find the function $F^{-1}(x)$ satisfying

$$F\left(F^{-1}(x)\right) = 1 - \frac{\exp\left(-\frac{F^{-1}(x)}{\rho}\right)}{\left(1+F^{-1}(x)\right)^{M-1}} = x.$$

The following chain of implications hold.

$$\begin{aligned}
& F\left(F^{-1}(x)\right) && = x \\
\Leftrightarrow \quad & \left(\left(1+F^{-1}(x)\right)\exp\left(\tfrac{1+F^{-1}(x)}{(M-1)\rho}\right)\right)^{1-M} && = \exp\left(-\tfrac{1}{\rho}\right)(1-x) \\
\Leftrightarrow \quad & \tfrac{1+F^{-1}(x)}{(M-1)\rho}\exp\left(\tfrac{1+F^{-1}(x)}{(M-1)\rho}\right) && = \tfrac{1}{(M-1)\rho}\left(\exp\left(-\tfrac{1}{\rho}\right)(1-x)\right)^{\frac{1}{1-M}} \\
\Leftrightarrow \quad & F^{-1}(x) && = -1+(M-1)\rho W\left(\tfrac{\exp\left(\frac{1}{(M-1)\rho}\right)}{(M-1)\rho}(1-x)^{\frac{1}{1-M}}\right),
\end{aligned}$$

which completes the proof.

4.8.6 Proof of Theorem 4.4

By Theorem 4.3, it is enough to show that $f'\left(F^{-1}(x)\right) \le -\frac{f\left(F^{-1}(x)\right)}{(1+F^{-1}(x))}$ for all $x \in [0,1]$. To this end, let

$$g(x) = 1 + \frac{\left(1+F^{-1}(x)\right) f'\left(F^{-1}(x)\right)}{f\left(F^{-1}(x)\right)}. \tag{4.22}$$

To simplify $g(x)$ further, we first put $y = F^{-1}(x)$ in (4.22). Then

$$g(y) = 1 + \frac{1+y}{\frac{e^{-\frac{y}{\rho}}}{(y+1)^M}\left(\frac{1}{\rho}(y+1)+M-1\right)} \times \left(\frac{(1+y)^M e^{-\frac{y}{\rho}}\frac{1}{\rho} - (1+y)^M e^{-\frac{y}{\rho}}\left(\frac{1}{\rho}(y+1)+M-1\right)\frac{1}{\rho} - M(1+y)^{M-1}\left(\frac{1}{\rho}(y+1)+M-1\right)e^{-\frac{y}{\rho}}}{(1+y)^{2M}}\right).$$

After some further simplifications, we get

$$g(y) = 1 + \frac{\frac{1}{\rho}(y+1)}{\left(\frac{1}{\rho}(y+1)+M-1\right)} - \frac{1}{\rho}(y+1) - M.$$

Using Lemma 4.8, we can write y as $y = -1 + (M-1)\rho\bar{W}(x)$, where

$$\bar{W}(x) = W\left(\frac{\exp\left(\frac{1}{(M-1)\rho}\right)}{(M-1)\rho}(1-x)^{\frac{1}{1-M}}\right).$$

Hence, $g(x)$ can be given as

$$\begin{aligned} g(x) &= 1 + \frac{\bar{W}(x)}{\bar{W}(x)+1} - (M-1)\bar{W}(x) - M \\ &= -\frac{(M-1)\bar{W}(x)^2 + (2M-3)\bar{W}(x) + M - 1}{\bar{W}(x)+1}, \end{aligned}$$

which is always strictly negative for $M \geq 2$. This implies

$$f'\left(F^{-1}(x)\right) \leq -\frac{f\left(F^{-1}(x)\right)}{(1+F^{-1}(x))}$$

for all $x \in [0,1]$ when $M \geq 2$, which completes the proof.

Chapter 5
The Feedback Capacity Tradeoff

Opportunistic beamforming is a reduced feedback communication strategy for vector broadcast channels which requires partial channel state information (CSI) at the base station for its operation. Although reducing feedback, this strategy in its plain implementations displays a linear growth in the feedback load with the total number of users in the system n, which is an onerous requirement for large systems. This chapter focuses on a more stringent but realistic $O(1)$ feedback constraint on the feedback load. Starting with a set of statistically identical users, we obtain the tradeoff curve tracing the Pareto optimal boundary between feasible and infeasible feedback-capacity pairs for opportunistic beamforming. Any point on this tradeoff curve can be obtained by means of homogeneous decentralized threshold feedback policies, which are rate-wise optimal, in which a user feeds back only if the received signal quality is good enough. The chapter includes the derivation of these optimum policies, and further shows to what extent the $O(1)$ feedback constraint must be relaxed to achieve the same sum-rate scaling as with perfect CSI. Extensions of these results to heterogeneous communication environments in which different users experience non-identical path-loss gains are also provided. We also show how threshold feedback policies can be used to provide fairness in a heterogeneous system, while simultaneously achieving optimal capacity scaling. Although most of our results are asymptotic in the sense that they are derived by letting n grow large, they provide promising performance figures with a close match to the asymptotically optimal results when used in finite size systems.

5.1 Introduction

THE full utilization of the benefits of MIMO in a communication environment consisting of a multitude of mobile users (MUs) requires the base station to have some knowledge about the channel state information (CSI) of all MUs, either perfect [11,94,96], or partial [79,97]. This can be a very onerous capacity requirement on the uplink when compared with the gains achieved in the downlink. This chapter focuses on the impor-

tant issue of feedback-capacity tradeoff in multi-antenna systems, which can be a performance limiting criterion for future 4G wireless networks. Here, we show that we can limit the feedback requirements on the uplink to practical levels, yet still achieve near-perfect aggregate communication rate scaling in the downlink. We focus on the classical opportunistic beamforming framework which operates on partial CSI, and still achieves the optimum capacity scaling in single-cell MIMO systems [79], as an alternative to schemes requiring perfect CSI to maximize data communication rates. The downlink throughput scales like $M \log\log n$, where M is the number of transmit antennas at the base station, and n is the number of MUs in the system. Although opportunistic beamforming reduces feedback considerably in comparison to having full CSI in its plain implementations, the number of MUs feeding back still grows linearly with the total number of MUs in the system. As a result, to achieve modest double-logarithmic multiuser diversity gains at the downlink through opportunistic scheduling, we still need to improve the capacity of the feedback link linearly with the number of MUs in the system, which is an impractical feedback requirement on the feedback channel. A solution which alleviates this impracticality is selective feedback, which we studied extensively in Chapter 3 and 4, where only a small subset of MUs (*i.e.*, MUs having the best instantaneous channel states) are multiplexed on the uplink feedback channel. Since the base station can communicate only with up to M MUs, simultaneously, an ideal selective feedback policy should allow only M MUs to feed back, *i.e.*, $O(1)$ feedback requirement.

5.1.1 Contributions

Firstly, we will use the results in Chapter 3 and 4 to obtain the structure of the downlink sum-rate maximizing selective decentralized feedback policy. We show that the sum-rate maximizing selective decentralized feedback policy is a *homogeneous* threshold feedback policy, where each MU requests a beam if its SINR value on that particular beam is above a given threshold value.

Then, we show how to set threshold values to ensure $O(1)$ feedback load on the average as the number of MUs in the system grows large. We illustrate the sub-optimal and optimal throughput scaling of such a system through the feedback-capacity trade-

off curve tracing the Pareto optimal boundary between feasible and infeasible feedback-capacity pairs. If a feedback-capacity pair (λ, c) lies above this curve, then it is an infeasible operating point (*i.e.*, there is no feedback policy achieving the capacity scaling of $c \log\log n$ subject to the restriction that no more than λ MUs feed back on the average), and if it is below this curve, then it is a feasible operating point (*i.e.*, there is a feedback policy achieving the capacity scaling of $c \log\log n$ without violating the feedback constraint λ). Furthermore, if a feedback policy operates at a point strictly below the tradeoff curve, then it is suboptimal in the sense that there is another feedback policy achieving the same capacity scaling with strictly less feedback. Our derived homogeneous threshold feedback policy achieves any point on this tradeoff curve. In particular, we show that there exists a sequence of homogeneous threshold feedback policies with appropriately chosen thresholds achieving any point on this curve. We also show that $M \log\log n$ scaling in [79] can be achieved by only allowing $O\left((\log n)^{\epsilon}\right)$ MUs to feed back on the average for any $\epsilon \in (0, 1)$.

The above contributions are for a system with statistically identical MUs. However, this assumption hardly holds in a practical system due to different path-loss gains experienced by the MUs in a cell. Obviously, the same threshold level for all MUs will not work for such a system. To be more specific, if the threshold is low, the MUs located close to the base station will feed back with very high probability. On the other hand, the MUs located far away from the base station will feed back with very low probabilities if the threshold level is set to a high value. In this work, we show that the threshold levels at different MUs can be systematically altered according to the user locations without violating the $O(1)$ feedback constraint. We obtain the feedback-capacity tradeoff curve for this system, and we again show that relaxing the $O(1)$ feedback requirement to $O\left((\log n)^{\epsilon}\right)$ for any $\epsilon \in (0, 1)$ will be enough to achieve $M \log\log n$ scaling in [79].

Although providing optimum performance in throughput scaling under finite feedback constraints, providing fairness among MUs becomes an important issue for heterogeneous communication environments. If a beam is allocated to the MU having the best SINR, a MU situated far from the base station has a lower probability of being scheduled and may starve for data compared to a MU staying close to the base station. To address

this issue, we introduce two new scheduling policies. These scheduling policies coupled up with the systematic alteration of the threshold levels at different MUs allow the system to achieve the optimum throughput scaling with the added advantage of ensuring fairness among MUs.

Most of our results in this chapter are asymptotic in the sense that they are derived by letting the number of MUs in the system grow large. Hence, we have also performed some numerical evaluations to illustrate the accuracy of our results for finite size systems. In particular, we show that the threshold levels set by using our asymptotic formulas clearly achieve $O(1)$ feedback in finite size systems as well. Although the closeness of the simulated results to the theoretical results reduce a little when M is increased, $O(1)$ constraint is well preserved, and shows a clear convergence to the theoretical result when the number of MUs is increased. We also compare the two proposed fair scheduling policies with the scheduling policy of allocating the beam to the best MU for finite n. Although achieving the same asymptotic performance, we observe that ensuring fairness causes a clear degradation in rate for any finite size system. This is, in fact, the tradeoff between rate maximization and ensuring fairness in a wireless communication network.

5.1.2 Related Works

The related work to this chapter includes most of the references given in Chapter 3 and 4. We study the opportunistic beamforming framework , but we have finite feedback constraints on the uplink in contrast to a feedback load which increases linearly with the number of MUs in [79, 97]. Quantization of the CSI parameters is an obvious way of further facilitating the feedback reduction for opportunistic beamforming, *e.g.*, see [39, 65, 78]. Although this technique helps to reduce the feedback considerably, it still cannot eliminate the linear growth in the feedback load. Therefore, opportunistic beamforming techniques were coupled with selective feedback techniques to further increase feedback performance [21,70,74,79]. Thresholding at the receiver, which has been proven to be the rate-wise optimal decentralized feedback policy in [74], has been the most commonly used selective feedback technique in the literature. The authors of [74] also prove the rate-wise optimality of homogeneous threshold feedback policies for a base station

transmitting multiple orthonormal beams in a Rayleigh fading environment. Our work focuses on homogeneous threshold feedback policies to obtain $O(1)$ feedback. Therefore, these results are directly related to the work in this chapter, and ensures zero optimality gap created by just focussing on the class of homogeneous threshold feedback policies for asymptotic throughput scaling. Different from [74], this chapter focuses on the optimum asymptotic throughput scaling under finite feedback constraints, and derives the feedback-capacity tradeoff curve.

A thresholding scheme for opportunistic beamforming was first proposed in [79], and then analyzed in greater detail in [70], but they only used a constant threshold independent of the number of MUs. Although such a constant thresholding scheme reduces the feedback, it cannot eliminate the linear growth in the average number of MUs feeding back, which renders them inappropriate for systems with large numbers of MUs but requiring only finitely many of them to feed back on the uplink. This discussion also implies that an ideal threshold should be a function of n. To this end, the results presented in [21] are promising to some extent where they showed that the average number of MUs feeding back can be reduced to $\log n$ by means of thresholding without any performance loss in the downlink throughput scaling. The main focus of our work is the further reduction of the average number of MUs feeding back to a more practical $O(1)$ level, and to investigate the capacity scaling laws with $O(1)$ feedback requirements. In this work, as a by-product of our analysis, we also show that the logarithmically growing feedback requirement can be further reduced to $O\left((\log n)^{\epsilon}\right)$ for $\epsilon \in (0,1)$ while achieving optimal downlink sum-rate scaling given in [79]. It is *almost as if* a constant feedback load is enough to maintain optimum sum-rate scaling but not exactly.

One of the main concerns associated with opportunistic beamforming is fairness. To address this issue, an algorithm called *proportionally fair algorithm* is introduced in [97]. This takes into consideration the ratio between the requested rate and the average throughput for a particular MU and ensures long-term fairness among MUs in terms of average data rates achieved by prioritizing a MU with a high ratio over the others. [54] extends this algorithm for multiuser scheduling. In [79], they claim that fairness can be obtained automatically as a byproduct of having multiple beams since a MU located near

the base station gets a high signal power as well as high interference from other beams. Our work introduces a new approach for achieving fairness in opportunistic beamforming. We show that setting different threshold levels at different MUs according to their locations, and fine tuning the scheduling policy, allows the system to obtain fairness among MUs.

Related work also includes the literature studying the applicability of opportunistic beamforming to finite networks or smaller sets of MUs [41, 54, 56, 102]. These works make subtle changes to the feedback and scheduling policies given in [97] and [79] to achieve better performance or robustness for smaller networks. In this work, we also verify the applicability of our asymptotic results to finite size systems. Although our motivation in this work is clearly different than the one in [41,54,56,102], it is a promising future research direction how to make use of the concepts in these works to enhance the applicability of our asymptotic results to finite size systems.

The rest of the chapter is organized as follows. In Section 5.2, we put forward the problem formulation. We review some fundamental properties of user SINRs and basic concepts from the extreme value theory in Section 5.3. Next, in Section 5.4, we present the main result of this chapter, which is the feedback-capacity tradeoff curve for opportunistic beamforming. We extend these results to statistically non-identical MUs in Section 5.5 and the fairness issues are investigated in Section 5.6. Section 5.7 contains the numerical evaluations that illustrate the applicability of our results to finite size systems. Section 5.8 concludes the chapter.

5.2 Problem Formulation

In this chapter, we continue studying the multi-antenna single cell vector broadcast channel model given in Subsection 2.2.1 specifically for a Rayleigh fading model. To recall, the base station communicates with n mobile users. The base station has N_t transmit antennas, and each MU is equipped with a single receive antenna. The channel gains between the receive antenna of the ith MU and the transmit antennas of the base station are given by $\boldsymbol{h}_i = (h_{1,i}, \ldots, h_{N_t,i})^\top$, where $h_{k,i}$ is the channel gain between the kth transmit antenna

at the base station and the receive antenna at the ith MU. We assume $h_{k,i}$, $k = 1, \ldots, N_t$ and $i = 1, \ldots, n$, are independent and identically distributed (i.i.d.) random variables drawn from a zero mean and unit variance *circularly-symmetric complex Gaussian* distribution $\mathcal{CN}(0, 1)$, similar to the model given in Subsection 4.6.1.

The base station transmits M data streams intended for M different MUs along the directions of M orthonormal beamforming vectors $\left\{\boldsymbol{b}_k = (b_{1,k}, \ldots, b_{N_t,k})^\top\right\}_{k=1}^{M}$. The symbols of the kth stream are represented by s_k. The signal received by the ith MU is equal to

$$Y_i = \sqrt{\rho} \sum_{k=1}^{M} \boldsymbol{h}_i^\top \boldsymbol{b}_k s_k + Z_i, \tag{5.1}$$

where Z_i is the unit power (complex) Gaussian background noise and ρ is the transmit power per beam. With these normalized parameter selections, ρ also signifies the SNR per beam.

$\gamma_{m,i}$ is the SINR at beam m at user i, and it is given by

$$\gamma_{m,i} = \frac{|\boldsymbol{h}_i^\top \boldsymbol{b}_m|^2}{\rho^{-1} + \sum_{k=1, k\neq m}^{M} |\boldsymbol{h}_i^\top \boldsymbol{b}_k|^2}. \tag{5.2}$$

$\boldsymbol{\gamma}_i \in \mathbb{R}_+^M$ represents the SINR vector at user i. The elements of $\boldsymbol{\gamma}_i$ are identically distributed with a common marginal distribution F. The M-by-n SINR matrix of the whole n-user communication system is denoted by $\boldsymbol{\Gamma} \in \mathbb{R}_+^{M\times n}$. If $\boldsymbol{\Gamma}$ is known by the base station, then the aggregate communication rate can be maximized by choosing the best MU with the highest SINR per beam. However, this requires an excessive amount of feedback and information exchange between the base station and MUs. Here, we are interested in the rate maximization under finite feedback constraints.

The MUs feedback using a given feedback policy $\boldsymbol{\mathcal{F}}$. A definition of $\boldsymbol{\mathcal{F}}$ is given in Definition 3.1. We give a more refine definition which is more relevant to the work in this chapter below, where $\mathcal{M} = \{1, \ldots, M\}$, and $\mathcal{N} = \{1, \ldots, n\}$.

Definition 5.1. *A feedback policy* $\boldsymbol{\mathcal{F}} : \mathbb{R}_+^{M\times n} \mapsto \{\Omega \bigcup \{\varnothing\}\}^n$ *is an* $\{\Omega \bigcup \{\varnothing\}\}^n$*-valued function*

$$\boldsymbol{\mathcal{F}} = (\mathcal{F}_1, \ldots, \mathcal{F}_n)^\top, \tag{5.3}$$

where $\mathcal{F}_i : \mathbb{R}_+^{M\times n} \mapsto \Omega \bigcup \{\text{ø}\}$ *is the feedback policy of MU* i, Ω *is the set of all possible feedback packets*[1] *and* ø *represents the no-feedback state. We call* $\boldsymbol{\mathcal{F}}$ *a decentralized feedback policy if* $\mathcal{F}_i$ *is only a function of* $\boldsymbol{\gamma}_i$ *for all* $i \in \mathcal{N}$. *We call it a homogeneous decentralized feedback policy if all MUs use the same feedback policy,* i.e., $\mathcal{F}_i = \mathcal{F}_j$ *for all* $i, j \in \mathcal{N}$.

A feedback policy determines two key performance measures of interest. Firstly, it determines the downlink ergodic sum-rate in an n-user system under the feedback policy $\boldsymbol{\mathcal{F}}$, which is given by

$$R_n\left(\boldsymbol{\mathcal{F}}\right) = \mathsf{E}_{\boldsymbol{\Gamma}}\left[\sum_{m=1}^{M} \log\left(1 + \max_{i\in\mathcal{G}_m(\boldsymbol{\mathcal{F}}(\boldsymbol{\Gamma}))} \gamma_{m,i}\right)\right]. \tag{5.4}$$

Note that this is the same definition given in (3.1), but we are introducing the subscript n since it is an variable that we alter in this analysis, unlike the analysis done in the previous chapters. Given a feedback policy $\boldsymbol{\mathcal{F}}$, we will use $R_n\left(\boldsymbol{\mathcal{F}}\right)$ to measure the performance of $\boldsymbol{\mathcal{F}}$ along the rate dimension. The next performance measure of interest is the average number of MUs feeding back $\Lambda(\boldsymbol{\mathcal{F}})$, which is used to measure the performance of $\boldsymbol{\mathcal{F}}$ along the feedback dimension. $\Lambda(\boldsymbol{\mathcal{F}})$ can be written as

$$\Lambda(\boldsymbol{\mathcal{F}}) = \sum_{i=1}^{n} \mathsf{Pr}\left\{\mathcal{F}_i\left(\boldsymbol{\Gamma}\right) \neq \text{ø}\right\}. \tag{5.5}$$

We define the feedback-capacity pre-log scaling factor $C^\star(\lambda)$ as

$$C^\star(\lambda) = \limsup_{n\to\infty} \sup_{\boldsymbol{\mathcal{F}}:\Lambda(\boldsymbol{\mathcal{F}})\leq\lambda} \frac{R_n(\boldsymbol{\mathcal{F}})}{M\log\log n}. \tag{5.6}$$

It is not hard to see that $C^\star(\lambda)$ is a monotone increasing function of λ with $C^\star(0) = 0$ (*i.e.*, beams are randomly allocated to MUs without using any CSI) and $\lim_{\lambda\to\infty} C^\star(\lambda) = 1$ (*i.e.*, all MUs feedback their maximum SINR [79]). In this chapter, we will determine $C^\star(\lambda)$ for intermediate values of λ in $(0, \infty)$, and call the resulting curve the feedback-capacity tradeoff curve as it signifies the best achievable capacity scaling given the feedback con-

[1] Many different feedback mechanisms can be considered. Feeding back all the SINR values, feeding back only the maximum SINR, feeding back the two largest SINRs, are just few of these approaches. The structure of the feedback packet will depend on the feedback approach. Ω is the set containing all of these possible feedback packets.

straint λ.

We say a sequence of feedback policies $\left\{\boldsymbol{\mathcal{F}}^{(n)}\right\}_{n=1}^{\infty}$ is *asymptotically λ-optimal* if, for a given feedback constraint λ, we have $\Lambda\left(\boldsymbol{\mathcal{F}}^{(n)}\right) \leq \lambda$ for all n large enough and

$$\lim_{n\to\infty} \frac{R_n\left(\boldsymbol{\mathcal{F}}^{(n)}\right)}{M \log\log n} = C^{\star}(\lambda). \tag{5.7}$$

In the rest of the chapter, we will not allow any message passing among MUs, and therefore focus on the distributed scenario. Hence, we will seek for asymptotically λ-optimality within the class of decentralized feedback policies.

We have already proved in Chapter 3 that a threshold feedback policy will be the downlink rate maximizing decentralized feedback policy when such a constraint on the average number of users feeding back is enforced (please refer Theorem 3.2 and Theorem 3.3). Therefore, we can only focus on the class of decentralized threshold feedback policies $\boldsymbol{\mathcal{T}}$ without any loss in optimality. Definitions for a threshold feedback policy is given in Definition 3.3 and 3.4. As we did for the definition of the feedback policy, we will conclude this section by giving a more refine definition for threshold feedback policies, which is an important subset of decentralized feedback policies since they become asymptotically λ-optimal for appropriately chosen thresholds as we will see in Section 5.4.

Definition 5.2. *We say $\boldsymbol{\mathcal{T}} = (\mathcal{T}_1, \ldots, \mathcal{T}_n)^{\top}$ is a threshold feedback policy if, for all $i \in \mathcal{N}$, there is a threshold τ_i such that $\mathcal{T}_i(\gamma_i)$ generates a feedback packet containing* SINR *values $\{\gamma_{k,i}\}_{k\in\mathcal{I}_i}$ if and only if $\gamma_{k,i} \geq \tau_i$ for all $k \in \mathcal{I}_i \subseteq \mathcal{M}$. We call it a homogeneous threshold feedback policy if all MUs use the same threshold τ,* i.e., *$\tau_i = \tau$ for all $i \in \mathcal{N}$.*

5.3 Signal-to-interference-plus-noise-ratio (SINR)

The SINR is an important parameter in our analysis since it is the metric that captures the quality of the channel, *i.e.*, see the rate definition in (5.4). Therefore, we will briefly review some of its fundamental properties in this section along with some basic concepts from the extreme value theory [19].

The SINR expression given in (5.2) for this specific system model consists of i.i.d. unit mean exponential random variables $\left\{|\boldsymbol{h}_i^\top \boldsymbol{q}_k|^2\right\}_{k=1}^{M}$ in the numerator and the denominator. Using this property, we can obtain the cumulative distribution function and the probability density function

Lemma 5.1. *The cumulative distribution function (CDF) and the probability density function (PDF) of the* SINR *on a beam at a user for the model in consideration are given by,*

$$F(x) = 1 - \frac{e^{-\frac{x}{\rho}}}{(x+1)^{M-1}} \tag{5.8}$$

and

$$f(x) = \frac{e^{-\frac{x}{\rho}}}{(x+1)^{M}} \left[\frac{1}{\rho}(x+1) + M - 1\right], \tag{5.9}$$

respectively.

Proof. See Appendix 5.9.1. □

Using these two functions and the result in Theorem 4.4, we can obtain the rate-wise optimality of homogeneous threshold feedback policies. This optimality ensures zero optimality gap, which can be possibly created by just focussing on the class of homogeneous threshold feedback policies for asymptotically λ-optimality. Therefore, we will analyze,

$$C^\star(\lambda) = \limsup_{n\to\infty} \sup_{\boldsymbol{\mathcal{T}}:\Lambda(\boldsymbol{\mathcal{T}})\leq\lambda} \frac{R_n(\boldsymbol{\mathcal{T}})}{M \log\log n}, \tag{5.10}$$

where $\boldsymbol{\mathcal{T}}$ is a homogenous threshold feedback policy.

For a sequence of i.i.d. random variables $X_1, X_2, \cdots, X_n$ with the common cumulative distribution function (CDF) G, we define

$$X_n^\star = \max_{1\leq i\leq n} X_i. \tag{5.11}$$

The CDF of $X_n^\star$ is equal to G^n, and if there exists a sequence of constants a_n and b_n such that $\frac{X_n^\star - b_n}{a_n}$ converges in distribution as n goes to infinity, then $G^n\left(a_n x + b_n\right)$ converges to one of the three well-known distributions: *Frechet*, *Weibull*, or *Gumbel* [19], given by

i) Frechet $G_1(x;\alpha) = \begin{cases} 0 & x \leq 0, \alpha > 0 \\ \exp\left(-x^{-\alpha}\right) & x > 0 \end{cases}$

ii) Weibull $G_2(x;\alpha) = \begin{cases} \exp\left[-(-x)^{\alpha}\right] & x \leq 0, \alpha > 0 \\ 1 & x > 0 \end{cases}$

iii) Gumbel $G_3(x) = \left\{ \exp\left[-e^{-x}\right] \quad -\infty < x < \infty \right.$

It is also known that the limiting distribution is of type *iii*, if the growth function, which is defines by

$$g(x) = \frac{1 - F(x)}{f(x)},$$

satisfies the von Mises' sufficient condition. It is not hard to see that

$$\lim_{x\to\infty} \frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{1-F(x)}{f(x)}\right] = \lim_{x\to\infty} \frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{(1+x)}{\left[\frac{1}{\rho}(x+1)+M-1\right]}\right] = 0, \tag{5.12}$$

which indicates that the growth function satisfies the von Mises sufficient condition. This implies that if

$$\gamma_m^{\star}(n) = \max_{1\leq i\leq n} \gamma_{m,i},$$

there exist sequences of real numbers $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ such that $\frac{\gamma_m^{\star}(n)-b_n}{a_n}$ converges in distribution to the Gumbel distribution, *i.e.*, for any $x \in \mathbb{R}$,

$$F^n\left(a_n x + b_n\right) \to \exp\left(-\mathrm{e}^{-x}\right)$$

as n goes to infinity. We formally state the sequences of real numbers a_n and b_n in the following lemma.

Lemma 5.2. *The sequences of real numbers $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ for the system model in consideration are given by,*

$$a_n = \frac{\rho(1+b_n)}{1+b_n+\rho(M-1)} \tag{5.13}$$

and

$$b_n = \rho \log n - \rho(M-1) \log\left(1 + b_n\right), \tag{5.14}$$

respectively.

Proof. See Appendix 5.9.2. □

In the next lemma, we establish an important structural property for the sum-rate function $R_n\left(\boldsymbol{\mathcal{T}}\right)$ when a homogeneous threshold feedback policy $\boldsymbol{\mathcal{T}}$ with a threshold level τ greater than 1 is used for MU selection. This property will simplify our analysis greatly in the remainder of the chapter.

Lemma 5.3. *Let $\boldsymbol{\mathcal{T}}$ be a homogeneous feedback policy with threshold $\tau > 1$. Then,*

$$R_n\left(\boldsymbol{\mathcal{T}}\right) = M\left(\mathsf{E}_{\boldsymbol{\Gamma}}\left[\log\left(1+\gamma_1^\star(n)\right) 1_{\{\mathcal{G}_1(\boldsymbol{\mathcal{T}}(\boldsymbol{\Gamma}))\neq\emptyset\}}\right]\right),$$

where $1_{\{\cdot\}}$ is the indicator function. Moreover, if a MU feeds back, it only requests the best beam with the maximum SINR.

Proof. See Appendix 5.9.3. □

An important theoretical consequence of this lemma is that it shows that it is enough to focus on the scaling properties of $\gamma_m^\star(n)$ to derive the scaling behavior of the sum-rate function. As will be seen in Section 5.4, scaling behavior of $\gamma_m^\star(n)$ can be obtained conveniently by using the extreme value theory. Moreover, since one of the main objectives of this chapter is to eliminate the linear growth in feedback load with the number of MUs, an assumption of a threshold being larger than one is not actually a limitation. That is, to eliminate the linear growth, the common threshold at MU terminals should gradually increase with the number of MUs. Therefore, if n is large enough, there exists $n^\star$ such that for all $n > n^\star$, $\tau > 1$. Furthermore, note that the lower bound on the threshold is 0[dB], which is a realistic figure even in a practical system with finite number of MUs.

5.4 Feedback-Capacity Tradeoff and Threshold Feedback Policies

In this section, we will prove the asymptotic λ-optimality of homogeneous threshold feedback policies with appropriately chosen threshold levels, and obtain the feedback-capacity tradeoff curve $C^{\star}(\lambda)$. One natural way of feedback reduction by means of a threshold feedback policy is to use a constant threshold τ, which is independent of n. Even though this approach reduces the number of MUs feeding back per beam by a factor of $1 - F(\tau)$, it cannot eliminate the linear feedback growth measured in terms of the number of MUs feeding back. Therefore, more stringent but realistic $O(1)$ feedback constraints on the feedback channel render constant thresholding approach useless for all practical purposes. As a result, we need threshold levels to grow to infinity to achieve $O(1)$ feedback as the number of MUs in the system increases.

In the next part of the chapter, roughly speaking, we will show that we can make a sequence of thresholding rules $\left\{\boldsymbol{\mathcal{T}}^{(n)}\right\}_{n=1}^{\infty}$ asymptotically λ-optimal by setting the threshold level of $\boldsymbol{\mathcal{T}}^{(n)}$ to

$$\tau_n(x) = a_n x + b_n, \tag{5.15}$$

where a_n and b_n are sequences of constants given in (5.13) and (5.14), respectively. Here, x is our design degree of freedom that we use to slide on the tradeoff curve $C^{\star}(\lambda)$ to achieve optimal capacity scaling without violating the feedback constraint.

5.4.1 Feedback-Capacity Tradeoff Curve

In Theorem 5.1, we establish the form of the feedback-capacity tradeoff curve $C^{\star}(\lambda)$. Before going into the details of the proof of Theorem 5.1, we will first provide some intuition for the proof. Consider a threshold feedback policy $\boldsymbol{\mathcal{T}}^{(n)}$ with a threshold level chosen as $\tau_n(x) = a_n x + b_n$.[2] Then, the random number of MUs requesting beam 1 when

[2]The form of threshold feedback policies that we use to prove Theorem 5.1 will be slightly different for technical purposes.

there are n MUs in the system can be written as

$$\lambda_{n,1}(x) \quad = \quad \sum_{i=1}^{n} 1_{\{\gamma_{1,i} \geq a_n x + b_n\}} \tag{5.16}$$

for all n large enough (see the proof of Lemma 5.3). Note that $\lambda_{n,1}(x)$ is a binomial random variable with parameters

$$\theta_n(x) = 1 - F(a_n x + b_n)$$

and n, *i.e.*,

$$\lambda_{n,1}(x) \sim B\left(n, \theta_n(x)\right)$$

. Since $\gamma_1^\star(n)$ converges in distribution to the Gumbel distribution, we have

$$\lim_{n\to\infty} \left(1 - \theta_n(x)\right)^n = \exp\left(-\mathrm{e}^{-x}\right). \tag{5.17}$$

Using this convergence, we can obtain an important result, which is given in the following lemma.

Lemma 5.4. *For any fixed $x \in \mathbb{R}$, if $\lim_{n\to\infty} \left(1 - \theta_n(x)\right)^n = \exp\left(-\mathrm{e}^{-x}\right)$, then*

$$\lim_{n\to\infty} n\theta_n(x) = \mathrm{e}^{-x}. \tag{5.18}$$

Proof. See Appendix 5.9.4. □

By using Lemma 5.4 and invoking the classical Poisson approximation for Binomial distributions, we conclude that $\lambda_{n,1}(x)$ can be approximated in distribution by a Poisson random variable with mean $n\theta_n(x)$ for n large enough. Furthermore, Lemma 5.4 also shows the form threshold levels achieving the required $O(1)$ feedback constraints, *i.e.,*

$$\lim_{n\to\infty} \Lambda\left(\boldsymbol{\mathcal{T}}^{(n)}\right) = M\mathrm{e}^{-x}. \tag{5.19}$$

However, we now encounter with the problem of outage. Above arguments indicate that the outage event occurs with probability around $\exp\left(-\mathrm{e}^{-x}\right)$ for n large enough. For

example, if we set $x = 0$, there will be approximately M MUs feeding back on the average but each beam will face an outage event with probability around 0.37 for n large enough, which shows a tradeoff between feedback and the capacity loss due to outage. In the next theorem, we make these intuitive ideas rigorous.

Theorem 5.1. *The feedback-capacity tradeoff curve $C^\star(\lambda)$ of a single-cell MIMO communication system with M transmit antennas at the base station is given by*

$$C^\star(\lambda) = 1 - \exp\left(-\frac{\lambda}{M}\right), \tag{5.20}$$

where λ is the $O(1)$ feedback constraint.

Proof. See Appendix 5.9.5. □

Theorem 5.1 clearly indicates how much loss in the capacity we have to tolerate in order not to violate a given feedback constraint on the feedback channel. By relaxing the feedback constraint, we can decrease outage probability and achieve higher rates of communication. Our design parameter x helps us to adjust threshold levels so that we can slide on the feedback-capacity tradeoff curve in the right direction to maximize communication rates while meeting the feedback constraint.

Fig. 5.1 shows feedback-capacity tradeoff curves for different numbers of transmitter antennas at the base station. Any feedback-capacity pair (λ, c) below these curves can be achieved by means of a homogeneous decentralized feedback policy, whereas there is no such policy that can achieve (λ, c) above them. Moreover, if a homogeneous decentralized feedback policy operates at a point strictly below the tradeoff curves, then it is sub-optimal in the sense that we can find another homogeneous decentralized feedback policy achieving the same capacity scaling with strictly less feedback. Therefore, these curves can also be considered as the Pareto optimal boundary between the feedback constraint and capacity scaling in single-cell MIMO communication systems.

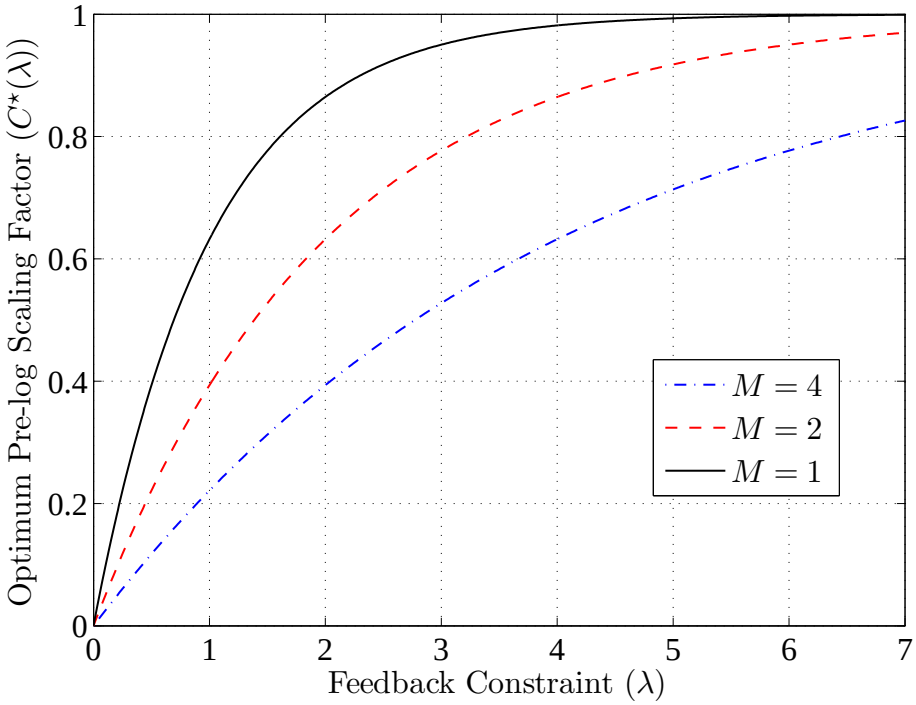


Figure 5.1: Feedback-capacity tradeoff curves for different numbers of transmitter antennas: $M = 1$, $M = 2$ and $M = 4$.

5.4.2 Closing the Capacity Gap with $O\left((\log n)^{\epsilon}\right)$ Feedback

As shown in the previous section, when there is a stringent $O(1)$ feedback constraint λ on the feedback channel, there will always be a capacity loss in the downlink of a single-cell MIMO communication system, albeit a small one for large but finite values of λ. In this part of the chapter, we will now show that we can even get rid of this capacity loss by sending the design parameter x, used to set threshold levels above, slowly to $-\infty$ if the $O(1)$ feedback constraint is relaxed to $O\left((\log n)^{\epsilon}\right)$ for any $\epsilon > 0$.

To this end, we consider a sequence of homogeneous decentralized threshold feedback policies $\left\{\boldsymbol{\mathcal{T}}^{(n)}\right\}_{n=1}^{\infty}$ with threshold levels

$$\tau_n(x_n) = a_n x_n + b_n \tag{5.21}$$

such that

$$x_n = -\epsilon \log\log n \tag{5.22}$$

for some $\epsilon > 0$. As above, let $\lambda_{n,1}(x_n)$ be the random number of MUs requesting beam 1 under the thresholding rule $\boldsymbol{\mathcal{T}}^{(n)}$ when there are n MUs in the system. Then, by using Lemma 5.3, we can write

$$\lambda_{n,1}(x_n) = \sum_{i=1}^{n} 1_{\{\gamma_{1,i} \geq a_n x_n + b_n\}}$$

for all n large enough. $\lambda_{n,1}(x_n)$ has a binomial distribution with parameters

$$\theta_n(x_n) = 1 - F\left(a_n x_n + b_n\right)$$

and n. Therefore, the outage probability at beam 1 is equal to

$$\beta_n = \left(1 - \theta_n(x_n)\right)^n$$

and

$$\Lambda\left(\boldsymbol{\mathcal{T}}^{(n)}\right) = Mn\theta(x_n).$$

We will now obtain an expression for $\theta_n\left(-\epsilon \log\log n\right)$, which is formally stated in the following lemma.

Lemma 5.5. *For some positive sequence of real numbers $\{k_n\}_{n=1}^{\infty}$ such that $k_n = O(1)$, we have*

$$\theta_n(-\epsilon \log\log n) = \frac{1}{n}\left(\log n\right)^{\epsilon} k_n \tag{5.23}$$

for the system in consideration.

Proof. See Appendix 5.9.6. □

Using this result, we have the average feedback around $O\left((\log n)^{\epsilon}\right)$. Furthermore, we can obtain an expression for the outage probability, which we obtain through the following lemma.

Lemma 5.6. *For these choices of threshold levels*, i.e., $\tau_n(x_n) = a_n x_n + b_n$, *where* $x_n = -\epsilon \log\log n$, *we have the outage probability*

$$\beta_n = O\left(\frac{1}{(\log n)^{\epsilon}}\right) \tag{5.24}$$

for the system in consideration.

Proof. See Appendix 5.9.7. □

Since the outage probability goes to zero, we do not have any optimality gap in the scaling of the system capacity. The following theorem formally characterizes this result.

Theorem 5.2. *A sequence of thresholding rules* $\left\{\boldsymbol{\mathcal{T}}^{(n)}\right\}_{n=1}^{\infty}$ *with thresholds* $\tau(x_n) = a_n x_n + b_n$, *where* a_n *and* b_n *are given in* (5.13) *and* (5.14), *respectively, and* $x_n = -\epsilon \log\log n$ *for any* $\epsilon \in (0,1)$, *achieves the* $M \log\log n$*-type capacity scaling behavior, i.e.,*

$$\lim_{n\to\infty} \frac{R_n\left(\boldsymbol{\mathcal{T}}^{(n)}\right)}{M \log\log n} = 1, \tag{5.25}$$

with feedback load growing as $\Lambda\left(\boldsymbol{\mathcal{T}}^{(n)}\right) = O\left((\log n)^{\epsilon}\right)$.

Theorem 5.2 can be proven by following the same lines in the proof of Theorem 5.1, and therefore we omit the details. One implication of this theorem is that we can achieve the optimal capacity scaling with an "almost" $O(1)$ feedback by setting ϵ close to zero. However, we cannot make ϵ exactly zero, and therefore the feedback increases to infinity but at a much slower rate than $\log n$ for small values of ϵ. The analysis in this part of the chapter extends similar findings in [21], where they only showed that $O(\log n)$ feedback, *i.e.*, $\epsilon = 1$, is enough to achieve optimal capacity scaling. Here, we show that we can make the growth rate of the feedback arbitrarily close to constant feedback by introducing another design parameter ϵ to choose threshold feedback policies.

Finally, since threshold levels are chosen to grow like $\log n$ in all the analysis above, it will be enough to assume that each MU only feeds back its user index and the index of the best beam. Any MU, not only the MU with the maximum SINR requesting a particular beam, with SINR above the threshold will achieve $\log\log n$-type scaling. This means that

the base station can select one MU randomly for every beam among the MUs requesting it without any loss in the asymptotic performance.

5.5 Heterogeneous Communication Environments

The analysis in the earlier sections of the chapter was done for a homogeneous network with all the MUs having equal average SNR ρ, but in a practical system, this assumption hardly holds because the MUs are scattered throughout the cell. To model this scenario, we consider the setup given in the next subsection.

5.5.1 System Setup

We assume that the set of MUs is divided into L groups ($n >> L$) according to their average SNR values. ρ_k represents the SNR of the kth group. For notational simplicity, we will assume that each group has $\bar{n} = \frac{n}{L}$ MUs, *i.e.*, equal number of MUs. For a given set of SNR values $\{\rho_1, \ldots, \rho_L\}$, let

$$\rho_{\max} = \max_{1 \le l \le L} \rho_l \tag{5.26}$$

and

$$\rho_{\min} = \min_{1 \le l \le L} \rho_l, \tag{5.27}$$

and we assume that $0 < \rho_{\min} \le \rho_{\max} < \infty$. Also, let

$$l^{\star} = \arg \max_{1 \le l \le L} \rho_l, \tag{5.28}$$

which is the index of the best group, and

$$l^{\flat} = \arg \min_{1 \le l \le L} \rho_l, \tag{5.29}$$

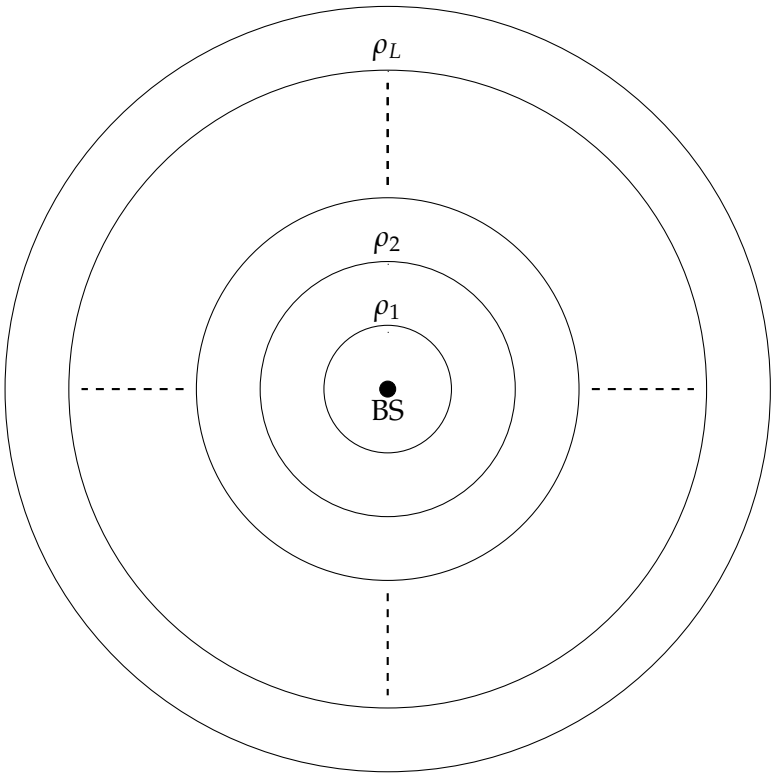


Figure 5.2: Setup for the analysis for heterogeneous communication environments.

which is the index of the worst group. Please refer to Fig. 5.2 for a possible graphical representation, where MUs whose distances to the base station in between some predefined limits are assigned to the same group, and are approximated as having the same average SNR value. Also, note that the MUs in a particular group, *e.g.,* let's say group k, are still statistically identical to each other with a common CDF for the SINR on a beam given by

$$F(x, \rho_k) = 1 - \frac{\mathrm{e}^{-\frac{x}{\rho_k}}}{(x+1)^{M-1}}. \tag{5.30}$$

Finally, let $\gamma^{\star}_m(\bar{n}, k)$ denote the maximum SINR on beam m among the $\bar{n}$ MUs in group k.

5.5.2 Throughput Scaling for a Heterogeneous System

In this part of the chapter, we will extend the results in Subsection 5.4.1 while taking the heterogeneity in SNR values into account. We will show that the heterogeneous network,

which is illustrated in Fig. 5.2, will achieve the same $C^\star(\lambda)$ scaling given in Theorem 5.1, for any given threshold levels achieving the feedback constraint λ. These ideas are formally given in the following theorem.

Theorem 5.3. *Consider a heterogeneous single-cell MIMO communication system with M transmit antennas at the base station where each MU has a bounded* SNR *from a finite set of* SNR *values. The feedback-capacity tradeoff curve $C^\star(\lambda)$ of this system is given by*

$$C^\star(\lambda) = 1 - \exp\left(-\frac{\lambda}{M}\right), \tag{5.31}$$

where λ is the $O(1)$ feedback constraint.

Proof. See Appendix 5.9.8. □

We can also achieve the $M \log\log n$-type throughput scaling behavior by setting x as a function of n and slowly decreasing it to $-\infty$ as we did in Subsection 5.4.2. The arguments for the proof and the analysis are similar to the fixed x case, and therefore we skip them to avoid repetition. The results are formally stated in the following theorem.

Theorem 5.4. *A sequence of thresholding rules $\left\{\boldsymbol{\mathcal{T}}^{(n)}\right\}_{n=1}^{\infty}$ with thresholds*

$$\tau_{\bar{n}}(x_{\bar{n}}, k) = a_{\bar{n}}(k)x_{\bar{n}} + b_{\bar{n}}(k),$$

where $k \in \{1, \ldots, L\}$ and $a_{\bar{n}}(k)$ and $b_{\bar{n}}(k)$ are given in (5.34) *and* (5.35)*, respectively, and $x_{\bar{n}} = -\epsilon \log\log \bar{n}$ for any $\epsilon \in (0,1)$, achieves the $M \log\log n$-type capacity scaling behavior, i.e.,*

$$\lim_{n\to\infty} \frac{R_n\left(\boldsymbol{\mathcal{T}}^{(n)}\right)}{M \log\log n} = 1, \tag{5.32}$$

with feedback load growing like $\Lambda\left(\boldsymbol{\mathcal{T}}^{(n)}\right) = O\left((\log n)^\epsilon\right)$.

5.5.3 Drawbacks of the Scheduling Policy

The main drawback of this scheduling policy is its inability to provide fairness among different MUs. If the MUs were homogeneous, fairness would be achieved as an automatic byproduct of the homogeneity assumption. However, if a beam is allocated to the MU having the best SINR in a heterogeneous communication environment, the MUs with low average SNR values will starve for data, which eliminates fairness in the system. The fairness among MUs can be defined using two different view points. The first one is providing the same rate for all the MUs, which we call rate-wise fairness. The second one is providing all the MUs the equal share of the channel in time, which we call time-wise fairness. In the next section, we will analyze how threshold feedback policies can be used to achieve time-wise fairness in a heterogeneous network while achieving the optimum throughput scaling under finite feedback constraints.

5.6 Obtaining Time-wise Fairness Through Thresholds

The first challenge in achieving fairness in a heterogeneous network is the process of setting the threshold levels at different MUs. Clearly, we will run into two problems by setting the same threshold value for all MUs in such a system. Firstly, all the MUs in groups with high SNR values will feedback with high probabilities if the threshold is set to a low value. Secondly, the MUs in groups with low SNR values will feedback with small probabilities if the threshold is set to a high value. Therefore, it is imperative to set different threshold levels for different MU groups to make sure that all MUs have an opportunity to feed back regardless of their relative channel conditions. To this end, the threshold for the kth group $k \in \{1, \ldots, L\}$ is chosen as

$$\tau_{\bar{n}}(x, k) = a_{\bar{n}}(k)x + b_{\bar{n}}(k), \tag{5.33}$$

where

$$a_{\bar{n}}(k) \quad = \quad \frac{\rho_k(1 + b_{\bar{n}})}{1 + b_{\bar{n}} + \rho_k(M - 1)}, \tag{5.34}$$

$$b_{\bar{n}}(k) \quad = \quad \rho_k \log \bar{n} - \rho_k (M-1) \log\left(1 + b_{\bar{n}}\right), \tag{5.35}$$

and x is the design degree of freedom. The techniques used in Section 5.3 to obtain (5.13) and (5.14) can also be used to obtain (5.34) and (5.35), respectively. All the MUs in the kth group will use this homogeneous threshold. It is not hard to see that the threshold value is systematically reduced for MUs with poor average SNR values compared to MUs with high SNR. This ensures that all the MUs have an opportunity to feed back regardless of their relative channel conditions.

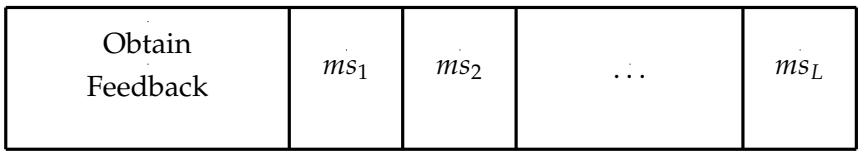
In the previous section, we claimed that the current scheduling scheme will not ensure time-wise fairness. We call this policy the *Best User Allocation Policy* (BUAP). Therefore, we will now define two new scheduling schemes which can be used to achieve fairness.

5.6.1 Time-wise Fair Scheduling Policies

Definition 5.3. Best User Fair Allocation Policy (BUFAP) *: The threshold for the kth group $k \in \{1, \ldots, L\}$ is chosen as $\tau_{\bar{n}}(x,k) = a_{\bar{n}}(k)x + b_{\bar{n}}(k)$ where x is the design degree of freedom, and $a_{\bar{n}}(k)$ and $b_{\bar{n}}(k)$ are given in* (5.34) *and* (5.35)*, respectively. Let $\mathcal{G}_{m,k}\left(\boldsymbol{\mathcal{T}}(\boldsymbol{\Gamma})\right)$ be the set of MUs feeding back on beam m from group k. Select the best MU $n^\star \in \mathcal{G}_{m,k}\left(\boldsymbol{\mathcal{T}}(\boldsymbol{\Gamma})\right)$ for each $m \in \mathcal{M}$ and each k. Allocate a mini-slot ms_k for each group $k \in \{1, \ldots, L\}$. All the mini-slots are equal in length in the time domain. Schedule communication to the selected MUs in each group on their allocated mini-slots and respective beams. Zero rate will be achieved on beam m in ms_k if $\mathcal{G}_{m,k}\left(\boldsymbol{\mathcal{T}}(\boldsymbol{\Gamma})\right)$ is an empty set, i.e., an outage event. Refer to Fig. 5.3(a) for a graphical representation of the time slot.*

Definition 5.4. Random User Allocation Policy (RUAP) *: The threshold for the kth group $k \in \{1, \ldots, L\}$ is chosen as $\tau_{\bar{n}}(x,k) = a_{\bar{n}}(k)x + b_{\bar{n}}(k)$ where x is the design degree of freedom, and $a_{\bar{n}}(k)$ and $b_{\bar{n}}(k)$ are given in* (5.34) *and* (5.35)*, respectively. Randomly select a MU $n' \in \mathcal{G}_m\left(\boldsymbol{\mathcal{T}}(\boldsymbol{\Gamma})\right)$ for each $m \in \mathcal{M}$ and schedule transmission to the selected MUs on the respective beams. Zero rate will be achieved on beam m if $\mathcal{G}_m\left(\boldsymbol{\mathcal{T}}(\boldsymbol{\Gamma})\right)$ is an empty set. Refer to Fig. 5.3(b) for a graphical representation of the time slot.*

When intuitively looking at the two scheduling policies, we can see that both of them



(a) Time slot for Best User Fair Allocation Policy (BUFAP).

(b) Time slot for Random User Allocation Policy (RUAP).

Figure 5.3: The two new scheduling policies for $M = 1$.

will provide some sense of fairness since they treat all the groups equally when scheduling communication on the downlink. Next, we will see how we can achieve fairness by using the scheduling policies defined above. Similar to our analysis in Section 5.4 and Section 5.5, we will start with a $O(1)$ feedback constraint and obtain the feedback-capacity tradeoff curve in a time-wise fair system. Then, we will see how we can set threshold levels in a time-wise fair system to ensure optimal throughput scaling.

5.6.2 Throughput Scaling for a Time-wise Fair System

We will now extend the results in Subsection 5.4.1 taking fairness considerations into account. Since these results are derived by allocating each beam to the MU having the maximum SINR on it, *i.e.*, the maximization among n random variables, we only use BUFAP given in Definition 5.3 for the theoretical analysis. Note that the best user allocation policy also allows a maximization in each of the mini-slots. Although there is no restriction in using the random user allocation policy for this scenario, the throughput scaling achieved with the random allocation policy will be strictly less for the fixed x scenario since the maximization among the MUs is neglected.

When x is fixed at a real number, we can write the random number of MUs requesting beam 1 as

$$\lambda_{n,1}(x) \quad = \quad \sum_{k=1}^{L} \bar{\lambda}_{\bar{n},1}(x,k)$$

$$= \sum_{k=1}^{L} \sum_{i \in \mathcal{S}_k, i=1}^{\bar{n}} 1_{\{\gamma_{1,i} \geq a_{\bar{n}}(k)x + b_{\bar{n}}(k)\}} \tag{5.36}$$

for all n large enough, where $\mathcal{S}_k$ is the group of MUs having ρ_k as their average SNR value. Again, $\bar{\lambda}_{\bar{n},1}(x,k)$ is a binomial random variable with parameters

$$\theta_{\bar{n}}(x,k) = 1 - F\left(a_{\bar{n}}(k)x + b_{\bar{n}}(k), \rho_k\right)$$

and $\bar{n}$, *i.e.*,

$$\bar{\lambda}_{\bar{n},1}(x,k) \sim B\left(\bar{n}, \theta_{\bar{n}}(x,k)\right).$$

$\lambda_{n,1}(x)$ is the summation of L independent binomial random variables. Since $\bar{n} \to \infty$ when $n \to \infty$, we know $\gamma_1^\star(\bar{n},k)$ converges in distribution to the Gumbel distribution for all k. This means, we have e^{-x} MUs feeding back from each group on a beam with an outage of $\exp\left(-\mathrm{e}^{-x}\right)$. Therefore, we have $O(1)$ feedback with

$$\lim_{n\to\infty} \Lambda\left(\boldsymbol{\mathcal{T}}^{(n)}\right) = ML\mathrm{e}^{-x}. \tag{5.37}$$

This also means that all the groups have the same feedback-capacity tradeoff curve. Since the BUFAP is used, time-wise fairness is guaranteed because an equal length mini-slot is allocated to each group for downlink data transmission. Hence, $O(1)$ feedback is achieved while preserving fairness. These ideas and the feedback-capacity tradeoff curve are formally given in the following theorem.

Theorem 5.5. *Consider a heterogeneous single-cell MIMO communication system with M transmit antennas at the base station where each MU has one of L bounded* SNR *values and the system ensures time-wise fairness using a the Best User Fair Allocation Policy . The feedback-capacity tradeoff curve $C^\star(\lambda)$ of this system is given by*

$$C^\star(\lambda) = 1 - \exp\left(-\frac{\lambda}{ML}\right), \tag{5.38}$$

where λ is the $O(1)$ feedback constraint.

Proof. Since the system achieves the same throughput scaling in all the mini-slots, the

theorem can be proven by considering the throughput scaling of a single mini-slot and following the lines of the proof of Theorem 5.1. □

We can again close this capacity gap with $O\left((\log \bar{n})^{\epsilon}\right)$ feedback by setting $x_{\bar{n}} = -\epsilon \log\log \bar{n}$ in a time-wise fair system as well. Now, $O\left((\log \bar{n})^{\epsilon}\right)$ MUs will feed back from each group, and the outage probability will go to zero as $n \to \infty$ since $\bar{n} = \frac{n}{L}$. We have also claimed in Subsection 5.4.2 that allocating the beam to the best MU feeding back or any random MU feeding back provides the same asymptotic performance in throughput scaling when x_n is varied. Therefore, the scheduling policies given in Definition 5.3 and Definition 5.4 can be both used to achieve fairness for this scenario. We formally state these results in the next theorem.

Theorem 5.6. *A sequence of thresholding rules* $\left\{\boldsymbol{\mathcal{T}}^{(n)}\right\}_{n=1}^{\infty}$ *with thresholds*

$$\tau_{\bar{n}}(x_{\bar{n}}, k) = a_{\bar{n}}(k)x + b_{\bar{n}}(k),$$

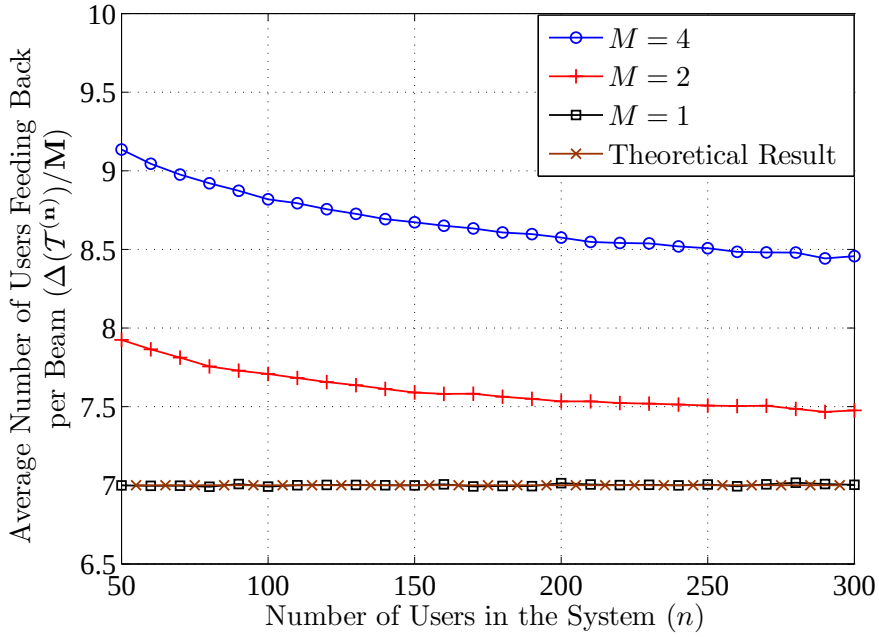
where $k \in \{1, \ldots, L\}$ *and* $a_{\bar{n}}(k)$ *and* $b_{\bar{n}}(k)$ *are given in* (5.34) *and* (5.35)*, respectively, and* $x_{\bar{n}} = -\epsilon \log\log \bar{n}$ *for any* $\epsilon \in (0,1)$*, achieves the* $M \log\log n$*-type capacity scaling, i.e.,*

$$\lim_{n\to\infty} \frac{R_n\left(\boldsymbol{\mathcal{T}}^{(n)}\right)}{M \log\log n} = 1, \tag{5.39}$$

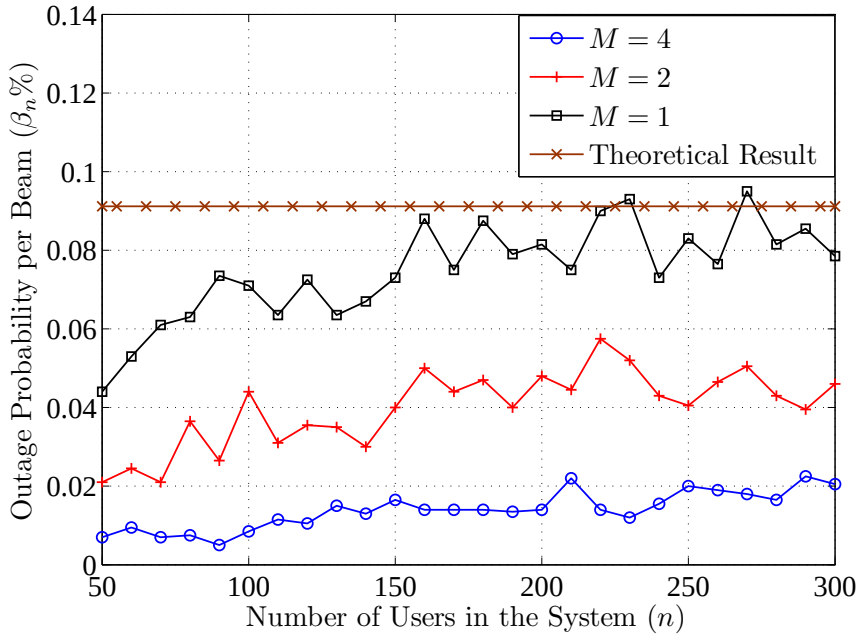
with feedback load growing as $\Lambda\left(\boldsymbol{\mathcal{T}}^{(n)}\right) = O\left((\log n)^{\epsilon}\right)$ *while ensuring time-wise fairness in the system.*

Proof. See Appendix 5.9.9. □

Theorems 5.5 and 5.6 show us that after some modifications to the scheduling policy, the results on throughput scaling hold for a heterogeneous set of MUs as well, and the system can simultaneously achieve time-wise fairness by fine tuning the threshold levels at different MUs according to their average SNR values.



(a) Change of $\Delta(\mathcal{T}^{(n)})$ per beam with n.



(b) Change of β_n per beam with n.

Figure 5.4: Simulation results illustrating the applicability of results in Subsection 5.4.1 to finite size systems, where $\rho = 1$.

5.7 Applicability to Finite Size Systems

Our analysis in the previous sections were based on asymptotic results. We will now see that some of our results provide good performance matches even for finite size networks. We start with the results obtained for the tradeoff curve in Subsection 5.4.1 and simulate the average number of MUs feeding back and the outage probability for various numbers of MUs in the range of 50 to 300 for $\rho = 1$.

The simulation results are illustrated in Fig. 5.4. We set $x = -\log 7$ which makes the average number of MUs feeding back 7 when $n \to \infty$. The simulated performance figures in Fig. 5.4(a) almost overlap with the theoretical results when $M = 1$, which makes the threshold setting ideal for finite number of MUs for this scenario. On the other hand, we can also observe that the tightness reduces a bit when M is increased. This can be reasoned out as follows. The simulated results will be close to the theoretical results if the distribution of the maximum SINR closely approximates the Gumbel distribution. When M is increased, interference is introduced, which causes the SINR to reduce. Hence, realized SINR values will achieve low values more frequently than high values. Since the distribution of the maximum SINR depends on high SINR values, which now have low probabilities of occurrence, we have to maximize over more random variables for the maximum SINR distribution to converge to the Gumbel distribution. This is evident in the Fig. 5.4(a) as well where we observe that the average number of MUs feeding back per beam converging to 7 as n is increased. When the value of ρ is decreased, *i.e.*, when the SINR is more noise dominated, the simulation results become tighter and the average number of MUs feeding back per beam is scattered closely around 7 even for the $M = 4$ case. It is also important to note that even when M is increased, we still have an upper bound on the average number of MUs feeding back per beam, which is approximately 9.15 for the $M = 4$ in Fig. 5.4(a). This implies that we can still upper bound the average number of MUs feeding back by using a constant λ, *i.e.*, $O(1)$ feedback load.

Outage probabilities are illustrated in Fig. 5.4(b). Theoretically, asymptotic outage probability should be around $\exp(-\exp(\log 7)) \approx 0.0009$. Again, our simulated results in the $M = 1$ case converge to this value rather quickly when compared to the simulation results in $M = 2$ and $M = 4$ cases. Also, we observe a decreasing behavior in β_n when

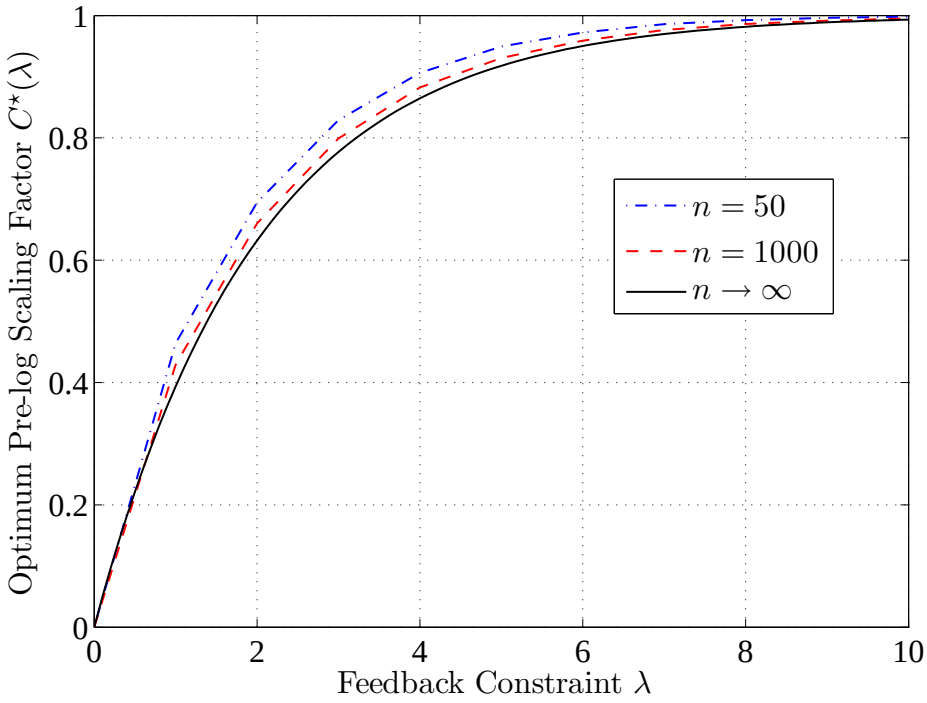
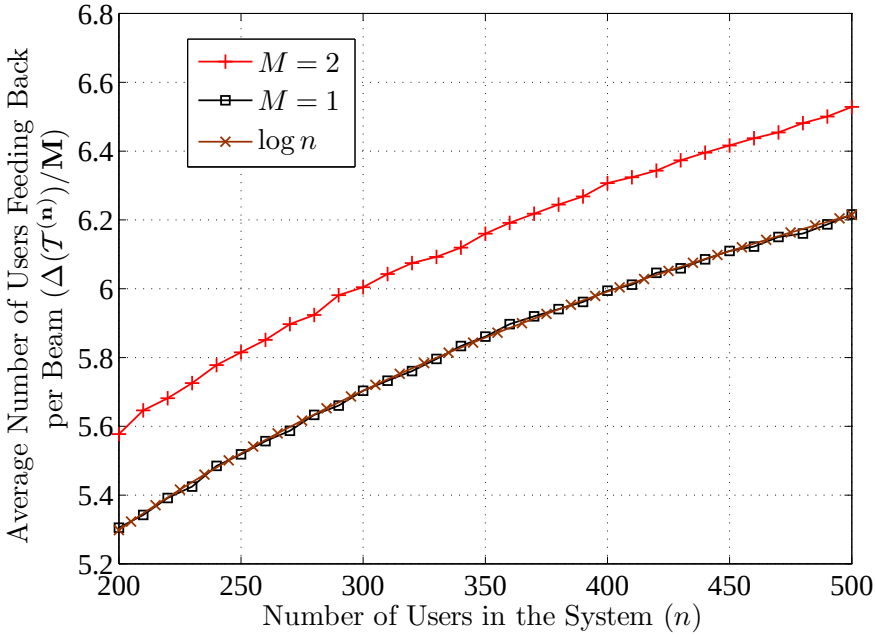


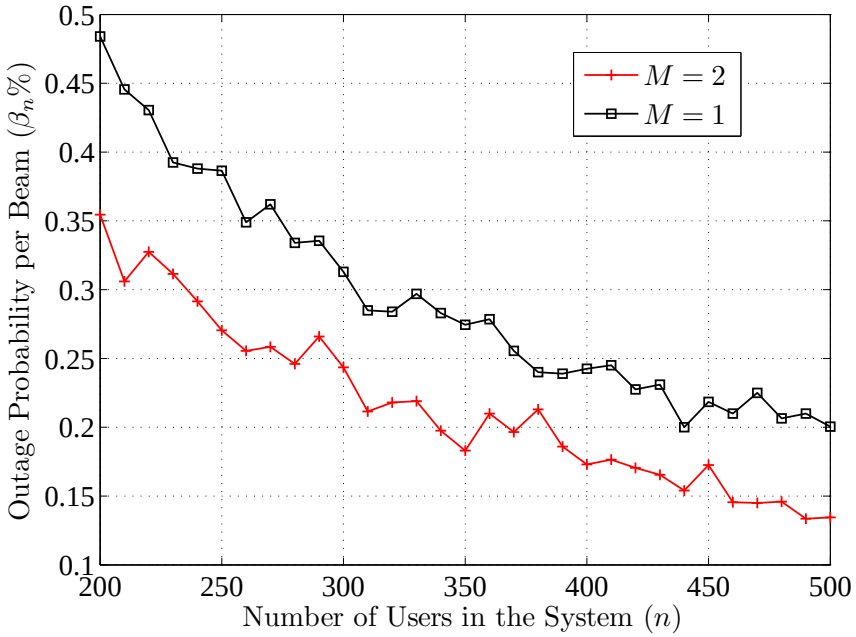
Figure 5.5: Convergence of the tradeoff curve as $n \to \infty$ for $M = 2$ and $\rho = 1$.

M is increased, which is consistent with Fig. 5.4(a). This is because increasing M causes more MUs to feed back for the finite n case. Fig. 5.5, which illustrates the convergence of the tradeoff curve as $n \to \infty$ for $M = 2$ and $\rho = 1$, concludes the numerical evaluations for the tradeoff curve. We observe that increasing the number of MUs causes $C^\star(\lambda)$ to decrease for any given feedback level λ (except for very small λ). This is because the outage probability increases with n as shown in Fig. 5.4(b).

Looking at the results in Subsection 5.4.2, we can show through simulations that the average number of MUs feeding back can be reduced from n to $O(\log n)$ with this selection of thresholds. We simulate the average number of MUs feeding back on a beam for different MU levels, where we set $\rho = 1$ and $\epsilon = 1$. The results are illustrated in Fig. 5.6(a). We observe that the average number of MUs feeding back can be reduced from n to $O(\log n)$ with the proposed selection of thresholds in Subsection 5.4.2. It is important to note that ϵ can be made smaller to further reduce the amount of feedback, but this makes the results more dependent on the magnitude of n since $\beta_n = O\left(\frac{1}{(\log n)^\epsilon}\right)$ and the



(a) Change of $\Delta(\mathcal{T}^{(n)})$ per beam with n.



(b) Change of β_n per beam with n.

Figure 5.6: Simulation results illustrating the applicability of results in Subsection 5.4.2 to finite size systems for $\rho = 1$ and $\epsilon = 1$.

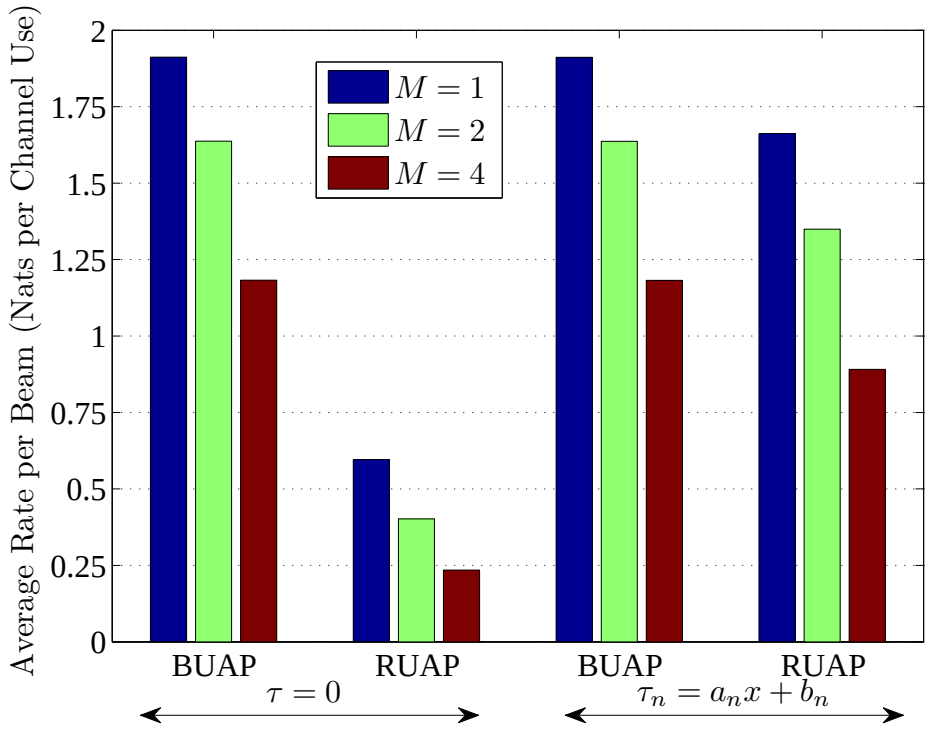


Figure 5.7: Simulation of rate for different scheduling policies and threshold selections for a set of homogeneous MUs, where $n = 200$ and $\rho = 1$.

rate at which β_n goes to zero reduces when ϵ is reduced. This is why we have set $\epsilon = 1$ for our simulations. When considering the outage probability, it is not difficult to see that the outage probability for any finite n will be again positive since β_n approaches zero only when $n \to \infty$. These results are illustrated in Fig. 5.6(b). As expected, the outage probability is high compared to the results obtained for the fixed x scenario, for finite n. However, the values for β_n are still relatively small, with it being less than 0.5% for the $M = 2$ case, and showing a clear convergence to zero as the number of MUs is increased.

Before focusing on the heterogeneous communication environments and the fairness issues, we will perform some more simulations with the assumption of homogeneity. We consider two cases, and simulate the rate for each case using different scheduling policies for $n = 200$ and $\rho = 1$. Firstly, we set the common threshold value to zero and allow all the MUs to feed back. The simulated rate is given in Fig. 5.7. Then, we set the common threshold level as $\tau_n = a_n x + b_n$, where $x = -\log 7$, and simulate the rate again. We observe that there is almost no degradation in rate on the average if BUAP is used but

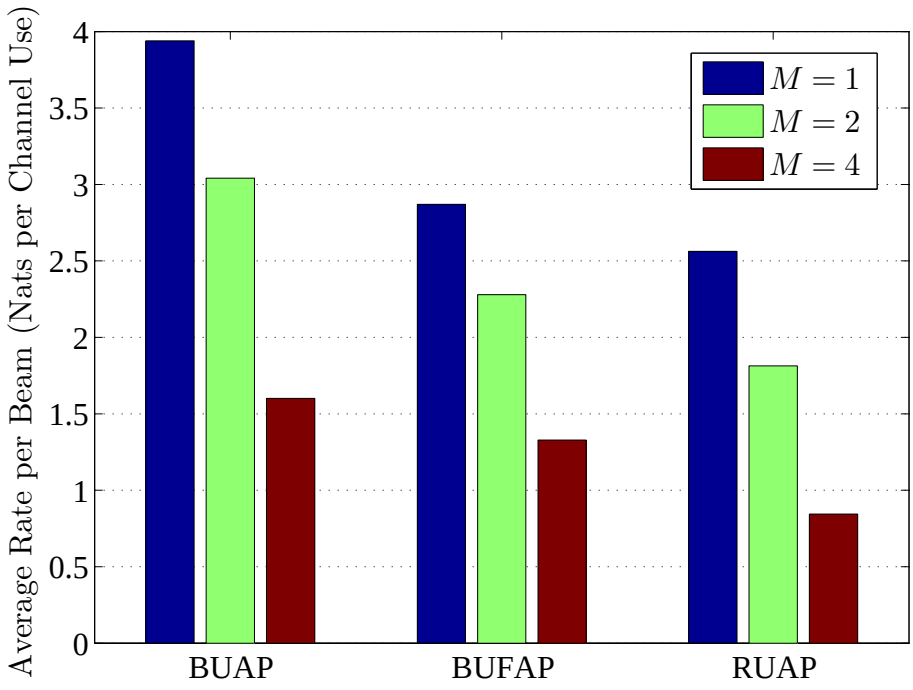


Figure 5.8: Simulation of rate for different scheduling policies for a heterogeneous network setting, where $n = 200$, $\rho_1 = 1$ and $\rho_2 = 10$.

the feedback load is reduced from 200 MUs to a figure approximately less than 8.5 MUs (please refer to Fig. 5.4(a) for the feedback levels at $n = 200$). On the other hand, setting a positive threshold provides a huge gain in terms of rate for RUAP since it stops a MU with low SINR from being scheduled for communication.

Now, we come to the scenario of heterogeneous MUs. We will assume two groups of MUs with each group having 100 MUs each. The first group has an average SNR $\rho_1 = 1$, and the second group has an average SNR $\rho_2 = 10$, *i.e.*, 0[dB] and 10[dB], respectively. Fig. 5.8 illustrates the simulated rate for this scenario with different scheduling policies. We can observe that BUAP does the best in terms of rate, and the other two scheduling policies achieve fairness at the expense of rate. This figure shows a clear tradeoff between rate maximization and ensuring fairness in a heterogeneous network setting. Finally, although both BUFAP and RUAP provide fairness, BUFAP is superior in terms of rate. However, if a RUAP is used, MUs have the added flexibility of not feeding back the SINR values to the base station. This implies a further reduction in feedback. It will also

simplify the feedback protocol implementation, which will be the focus of our future work.

5.8 Conclusions

Opportunistic beamforming is a communication strategy achieving the perfect CSI capacity scaling for vector broadcast channels by only requiring partial CSI at the base station. Nevertheless, in its plain implementations, it cannot eliminate the linear growth in the feedback load with increasing numbers of MUs in the system. In this chapter, we have focused on $O(1)$ feedback constraints on the feedback load instead, and analyzed the tradeoff between feedback and capacity in single-cell MIMO communication systems.

Starting with an assumption of statistically identical MUs, we have obtained the tradeoff curve $C^\star(\lambda)$ tracing the Pareto optimal boundary between feasible and infeasible feedback and capacity pairs. A point above $C^\star(\lambda)$ cannot be achieved by any decentralized feedback policy, whereas all points below it, which are sub-optimal in the sense that the same capacity scaling can be obtained by using strictly less feedback, can be achieved by such a policy. We have provided the form of the homogeneous threshold feedback policies, which are proven to be rate-wise optimal for Rayleigh fading environments with multiple transmit antennas at the base station, achieving the feedback-capacity pairs on $C^\star(\lambda)$. We have also showed that if the $O(1)$ feedback constraint λ is relaxed, we can achieve $M \log\log n$-type capacity scaling with feedback load growing like $O\left((\log n)^\epsilon\right)$ for any $\epsilon \in (0,1)$.

We have extended these results to a heterogeneous communication system, where different MUs experience non-identical path-loss gains. Using the same threshold level for all MUs does not work effectively for such a system. Therefore, we systematically altered the threshold levels at different MUs according to their large scale path-loss gains without violating the $O(1)$ feedback constraint. We have showed that the heterogeneous network achieves the same performance in capacity scaling and feedback as the network consisting of statistically identical MUs. Since providing fairness among MUs is an important issue for such a heterogeneous communication environment, we have introduced

two new scheduling policies. We have showed that these scheduling policies coupled up with the systematic alteration of the threshold levels at different MUs allow the system to achieve optimum throughput scaling and fairness among MUs, simultaneously.

Although most of our results are optimal in the limit for large networks, they also provide a close match with the asymptotically optimal results when used in finite size systems. In particular, we have showed that the threshold levels set by using the asymptotic formulas very rapidly achieve the required $O(1)$ feedback constraints in finite size systems as well. When comparing the two proposed fair scheduling policies with the scheduling policy of allocating the beam to the best MU, we have observed that ensuring fairness causes a clear degradation in rate for finite n although achieving the same asymptotic performance. This is the tradeoff between rate maximization and ensuring fairness in a wireless communication network.

5.9 Appendix

5.9.1 Proof of Lemma 5.1

First, let's consider the division of two independent random variables U and V. If $X = \frac{U}{V}$, the CDF of X can be obtained by

$$\mathsf{Pr}\{X \leq x\} = F(x) = \mathsf{Pr}\left\{\frac{U}{V} \leq x\right\},$$

which can be written as

$$\begin{aligned} F(x) &= \int_{v \in V} f_v(v).\mathsf{Pr}\left\{\frac{U}{v} \leq x\right\} dv \\ &= \int_{v \in V} f_v(v).\mathsf{Pr}\{U \leq vx\}\, dv, \end{aligned}$$

where f_v and f_u are the densities of V and U,respectively. If F_u is the CDF of U, we get

$$F(x) = \int_{v \in V} f_v(v) F_u(vx)\, dv. \tag{5.40}$$

The expression of the SINR given in (5.2) is the division of two random variables with a further added constant in the denominator:

$$\gamma_{m,i} = \frac{\overbrace{|\boldsymbol{h}_i^\top \boldsymbol{b}_m|^2}^{U}}{\underbrace{\frac{1}{\rho}}_{c} + \underbrace{\sum_{k=1,k\neq m}^{M} |\boldsymbol{h}_i^\top \boldsymbol{b}_k|^2}_{V}}.$$

Due to the orthogonality of the beams, we can assume independence between U and $c+V$. Therefore, we can make a slight change in (5.40) to take the constant c in to account. We get

$$F(x) = \int_{v\in V} f_v(v-c)F_u(vx)\,dv.$$

Also, $\left\{|\boldsymbol{h}_i^\top \boldsymbol{q}_k|^2\right\}_{k=1}^{M}$ are i.i.d. unit mean exponential random variables. Therefore, we have

$$F_u(vx) = 1 - e^{-vx}$$

and

$$f_v(v-c) = \frac{(v-c)^{(M-2)}e^{-(v-c)}}{(M-2)!}.$$

Substitution with a variable change $t = v - c$ gives us

$$F(x) = \frac{1}{(M-2)!}\int_0^\infty t^{(M-2)}e^{-t}\,dt - \frac{e^{-cx}}{(M-2)!}\int_0^\infty t^{(M-2)}e^{-t(1+x)}\,dt.$$

By using integral tables [32], and by substituting for c, we get

$$F(x) = 1 - \frac{e^{-\frac{x}{\rho}}}{(x+1)^{M-1}}. \tag{5.41}$$

Differentiating this expression with respect to x gives us the PDF $f(x)$.

5.9.2 Proof of Lemma 5.2

Firstly, according to the results in [19], if the extreme value distribution converges to the Gumbell distribution as n goes to infinity, the sequence of constants b_n can be found through the relation

$$1 - F(b_n) = \frac{1}{n}.$$

Substituting for F gives

$$\frac{e^{-\frac{b_n}{\rho}}}{(b_n+1)^{M-1}} = \frac{1}{n}.$$

which simplifies to

$$b_n = \rho \log n - \rho(M-1) \log\left(1 + b_n\right).$$

The other sequence of constants a_n, can be obtained using the relation

$$a_n = \frac{1 - F(b_n)}{f(b_n)},$$

which simplifies to

$$a_n = \frac{\rho(1+b_n)}{1 + b_n + \rho(M-1)}, \tag{5.42}$$

completing the proof.

5.9.3 Proof of Lemma 5.3

Since a common threshold is used for all the beams at a MU, and all the beams are statistically identical, it will be enough to focus on the first beam. The total rate will be equal to M times the rate achieved at the first beam.

Let $\mathcal{N}_1$ be the set of MUs having their best SINR value, which is greater than τ, at

beam 1, *i.e.*,

$$\mathcal{N}_1 = \{i : d_i^\star \geq \tau \quad \& \quad b_i^\star = 1\},$$

where $b_i^\star = \arg\max_{1\leq j\leq M} \gamma_{j,i}$ is the index of the beam with the maximum SINR at MU i, and $d_i^\star = \max_{1\leq j\leq M} \gamma_{j,i}$ is the value of this maximum SINR, for a given SINR matrix $\mathbf{\Gamma}$. We will first show that $\mathcal{G}_1(\boldsymbol{\mathcal{T}}(\mathbf{\Gamma})) = \mathcal{N}_1$. By definition of $\mathcal{G}_1(\boldsymbol{\mathcal{T}}(\mathbf{\Gamma}))$, we have

$$\mathcal{N}_1 \subseteq \mathcal{G}_1(\boldsymbol{\mathcal{T}}(\mathbf{\Gamma})).$$

To show the other direction, take any $i \in \mathcal{G}_1(\boldsymbol{\mathcal{T}}(\mathbf{\Gamma}))$, and a beam index $r \neq 1$. Then, $|\boldsymbol{h}_i^\top \boldsymbol{q}_1|^2 > |\boldsymbol{h}_i^\top \boldsymbol{q}_r|^2$ because $\tau > 1$. Therefore, the following chain of inequalities/equalities hold.

$$\begin{aligned}
\gamma_{1,i} &= \frac{|\boldsymbol{h}_i^\top \boldsymbol{q}_1|^2}{\frac{1}{\rho} + \sum_{k=2}^{M} |\boldsymbol{h}_i^\top \boldsymbol{q}_k|^2} \\
&> \frac{|\boldsymbol{h}_i^\top \boldsymbol{q}_r|^2}{\frac{1}{\rho} + \sum_{k=1,k\neq r}^{M} |\boldsymbol{h}_i^\top \boldsymbol{q}_k|^2} \\
&= \gamma_{r,i}.
\end{aligned}$$

Therefore, MU i achieves its maximum SINR, which is greater than τ, at beam 1, and belongs to $\mathcal{N}_1$. Moreover, all other SINR values are less than τ, and MU i only requests the best beam with the maximum SINR, which is beam 1. We finish the proof by observing that

$$\max_{i\in\mathcal{G}_1(\boldsymbol{\mathcal{T}}(\mathbf{\Gamma}))} \gamma_{1,i} = \gamma_1^\star(n)$$

since MUs who do not belong to $\mathcal{G}_1(\boldsymbol{\mathcal{T}}(\mathbf{\Gamma}))$ have SINR values less than τ.

5.9.4 Proof of Lemma 5.4

First of all, we note that $\theta_n(x)$ cannot be equal to 0 (or 1) infinitely many times, otherwise we can find a subsequence $\{n_k\}_{k=1}^\infty$ such that $\lim_{k\to\infty} (1-\theta_{n_k}(x))^{n_k} = 1$ (or 0), which implies x to be $+\infty$ (or $-\infty$), which is a contradiction. Therefore, without loss of generality,

we can assume $\theta_n(x) \in (0,1)$ for all n.

Next, we show that $\theta_n(x)$ converges to 0. To this end, assume that $c > 0$ is a limit point of $\theta_n(x)$. This implies that we can find a subsequence $\{n_k\}_{k=1}^{\infty}$ and $\epsilon \in (0,c)$ such that $\lim_{k\to\infty} \theta_{n_k}(x) = c$ and

$$\lim_{k\to\infty} (1-\theta_{n_k}(x))^{n_k} \leq \lim_{k\to\infty} (1-c+\epsilon)^{n_k} = 0,$$

which again creates a contradiction by implying x to be $-\infty$.

To finish the proof, we focus on $n\log(1-\theta_n(x))$. Since

$$\lim_{n\to\infty} (1-\theta_n(x))^n = \exp\left(-\mathrm{e}^{-x}\right),$$

we have

$$\lim_{n\to\infty} n\log(1-\theta_n(x)) = -\mathrm{e}^{-x}. \tag{5.43}$$

By using Taylor series expansion for $\log(1-\theta_n(x))$ around 1, we have

$$\begin{aligned} n\log(1-\theta_n(x)) &= -n\sum_{k=1}^{\infty} \frac{(\theta_n(x))^k}{k} \\ &= -n\theta_n(x)\left(1+\sum_{k=2}^{\infty} \frac{(\theta_n(x))^{k-1}}{k}\right) \\ &= -n\theta_n(x)\left(1+O\left(\theta_n(x)\right)\right). \end{aligned} \tag{5.44}$$

Since $\lim_{n\to\infty}\left(1+O\left(\theta_n(x)\right)\right) = 1$, (5.43) and (5.44) together imply $\lim_{n\to\infty} n\theta_n(x) = \mathrm{e}^{-x}$.

5.9.5 Proof of Theorem 5.1

It is enough to focus on the positive values of λ since it is easy to see that $C^\star(\lambda)$ is zero when there is no feedback. For a given feedback constraint $\lambda > 0$, we consider a sequence of threshold feedback policies $\left\{\boldsymbol{\mathcal{T}}^{(n)}\right\}_{n=1}^{\infty}$ with threshold levels

$$\tau_n(x+\epsilon_n) = a_n(x+\epsilon_n) + b_n,$$

where x is chosen to be $-\log\left(\frac{\lambda}{M}\right)$, a_n and b_n are given by (5.13) and (5.14), respectively, and ϵ_n is a sequence of real numbers going to zero, and is chosen to make $\Lambda\left(\boldsymbol{\mathcal{T}}^{(n)}\right) = \lambda$ for all n. It can be shown that such a sequence exists by using F, a_n and b_n given in (5.8), (5.13) and (5.14), respectively. Then, by assuming the limit below exists, which we will show it exists, we see that

$$\lim_{n\to\infty} \frac{R_n\left(\boldsymbol{\mathcal{T}}^{(n)}\right)}{M \log\log n} = C^\star(\lambda).$$

Therefore, it will be enough to analyze the limiting behavior of $\frac{R_n\left(\boldsymbol{\mathcal{T}}^{(n)}\right)}{M\log\log n}$ to obtain $C^\star(\lambda)$.

The lower bound on the limit behavior of $\frac{R_n\left(\boldsymbol{\mathcal{T}}^{(n)}\right)}{M\log\log n}$ is easy to obtain. When there is no outage, $R_n\left(\boldsymbol{\mathcal{T}}^{(n)}\right)$ becomes greater than $M\log\left(1 + a_n(x+\epsilon_n) + b_n\right)$. Let

$$\beta_n = \left(1 - \theta_n(x+\epsilon_n)\right)^n$$

be the outage probability at a particular beam, where $\theta_n(x)$ is as defined above. By using Slutsky's Theorem and putting $x = -\log\left(\frac{\lambda}{M}\right)$, it is easy to see that

$$\lim_{n\to\infty} \beta_n = \exp\left(-\frac{\lambda}{M}\right).$$

By using this result, we have

$$\liminf_{n\to\infty} \frac{R_n\left(\boldsymbol{\mathcal{T}}^{(n)}\right)}{M\log\left(1 + a_n(x+\epsilon_n) + b_n\right)} \geq 1 - \exp\left(-\mathrm{e}^{-x}\right).$$

Since $\lim_{n\to\infty} a_n = \rho$ and $b_n = O\left(\log n\right)$, we can equivalently write the above result as

$$\liminf_{n\to\infty} \frac{R_n\left(\boldsymbol{\mathcal{T}}^{(n)}\right)}{M\log\log n} \geq 1 - \exp\left(-\frac{\lambda}{M}\right).$$

To compute the upper bound, we analyze the tail probabilities lying under the distribution of $\gamma_1^\star(n)$. In particular, it can be shown that the following hold for $k \geq 1$:

$$\mathsf{Pr}\left\{\tau_n(x+\epsilon_n) \leq \gamma_1^\star(n) \leq a_n \log\log n + b_n\right\} = 1 - \beta_n - O\left(\frac{1}{\log n}\right),$$

$$\Pr\left\{a_n \log\log n + b_n < \gamma_1^\star(n) \le a_n \log n + b_n\right\} = O\left(\frac{1}{\log n}\right),$$
$$\Pr\left\{k a_n \log(n) + b_n < \gamma_1^\star(n) \le (k+1) a_n \log n + b_n\right\} = \frac{1}{n} O\left(\frac{\mathrm{e}^{O(k)}}{n^{k-1}}\right).$$

By using these estimates, we upper bound $R_n\left(\boldsymbol{\mathcal{T}}^{(n)}\right)$ as

$$R_n\left(\boldsymbol{\mathcal{T}}^{(n)}\right) \le M\left(1-\beta_n - O\left(\frac{1}{\log n}\right)\right)\log\left(a_n \log\log n + b_n\right) + O\left(\frac{\log\log n}{\log n}\right) + O\left(\frac{\log\log n}{n}\right).$$

Since $\lim_{n\to\infty} a_n = \rho$ and $b_n = O(\log n)$, we have

$$\limsup_{n\to\infty} \frac{R_n\left(\boldsymbol{\mathcal{T}}^{(n)}\right)}{M \log\log n} \le 1 - \exp\left(-\frac{\lambda}{M}\right),$$

which completes the proof.

5.9.6 Proof of Lemma 5.5

To start with, we have

$$\theta_n(x_n) = \frac{\exp - \left(\frac{a_n x_n + b_n}{\rho}\right)}{(a_n x_n + b_n + 1)^{M-1}} \tag{5.45}$$

using (5.8), where a_n and b_n are given in (5.13) and (5.14), respectively. From (5.14) we have

$$\begin{aligned}
b_n &= \rho \log n - \rho(M-1)\log\left(1 + \rho\log n - \rho(M-1)\log\left(1+b_n\right)\right) \\
&= \rho\log n - \rho(M-1)\log\left(\rho\log n\right) - \rho(M-1)\log\left(1 + \frac{1}{\rho\log n} - \frac{(M-1)\log(1+b_n)}{\log n}\right) \\
&= \rho\log n - \rho(M-1)\log\log n - \rho(M-1)\log\rho - \rho(M-1)\log\left(1 - O\left(\frac{\log\log n}{\log n}\right)\right) \\
&= \rho\log n - \rho(M-1)\log\log n - O(1).
\end{aligned}$$

By decreasing x_n slowly to $-\infty$ like $-\epsilon \log\log n$ for any $\epsilon \in (0,1)$, we have $\theta_n(-\epsilon \log\log n)$

$$\begin{aligned} &= \exp\left(\frac{a_n \epsilon \log\log n}{\rho} - \frac{b_n}{\rho} - (M-1)\log(1 - a_n \epsilon \log\log n + b_n)\right) \\ &= \exp\left(\frac{a_n \epsilon \log\log n}{\rho} - \log n + (M-1)\left(\log\log n - \log(1 - a_n \epsilon \log\log n + b_n)\right) + O(1)\right) \\ &= \exp\left(\frac{a_n \epsilon \log\log n}{\rho} - \log n + O(1)\right). \end{aligned}$$

Since $\lim_{n\to\infty} a_n = \rho$, we have, for some positive sequence of real numbers $\{k_n\}_{n=1}^{\infty}$ such that $k_n = O(1)$,

$$\theta_n(-\epsilon \log\log n) = \frac{1}{n}(\log n)^{\epsilon} k_n,$$

which completes the proof.

5.9.7 Proof of Lemma 5.6

To obtain the outage probability, we will first prove the following result. Let $f_n \to 0$ and $g_n \to \infty$ as $n \to \infty$. Then, $(1+f_n)^{g_n} \sim \exp(f_n g_n)$ if and only if $(f_n)^2 g_n \to 0$ as $n \to \infty$

First, we will show that if $(1+f_n)^{g_n} \sim \exp(f_n g_n)$, $(f_n)^2 g_n$ converges to zero as $n \to \infty$. If $(1+f_n)^{g_n} \sim \exp(f_n g_n)$, we have

$$\lim_{n\to\infty} \frac{\exp\left(g_n \log(1+f_n)\right)}{\exp(f_n g_n)} = 1 \quad \Rightarrow \quad \lim_{n\to\infty} \exp\left(f_n g_n \left(\frac{\log(1+f_n)}{f_n} - 1\right)\right) = 1$$

which means,

$$\lim_{n\to\infty} f_n g_n \left(\frac{\log(1+f_n)}{f_n} - 1\right) = 0.$$

Since f_n approaches zero as n goes to infinity, we can write

$$\frac{\log(1+f_n)}{f_n} - 1 = f_n \sum_{l=0}^{\infty} \frac{1}{l+2}(-1)^{l+1} f_n^l$$

using the power series expansion of $\log(1+f_n)$. As a result, we need to have

$$\lim_{n\to\infty} g_n f_n^2 \sum_{l=0}^{\infty} \frac{1}{l+2}(-1)^{l+1} f_n^l = 0.$$

Since f_n approaches zero as n goes to infinity, we have, for sufficiently large n,

$$\left| \frac{1}{l+2}(-1)^{l+1} f_n^l \right| \leq 0.5^l.$$

Then, by the Lebesgue dominated convergence theorem, we obtain

$$\begin{aligned} \lim_{n\to\infty} \sum_{l=0}^{\infty} \frac{1}{l+2}(-1)^{l+1} f_n^l &= \sum_{l=0}^{\infty} \lim_{n\to\infty} \frac{1}{l+2}(-1)^{l+1} f_n^l \\ &= \sum_{l=0}^{\infty} \frac{1}{l+2}(-1)^{l+1} 1_{\{l=0\}} = -\frac{1}{2}. \end{aligned}$$

Therefore, we conclude that $(f_n)^2 g_n$ converges to zero as n goes to infinity.

To show $(1+f_n)^{g_n} \sim \exp(f_n g_n)$ if $(f_n)^2 g_n \to 0$ as $n \to \infty$, observe that

$$\begin{aligned} \frac{(1+f_n)^{g_n}}{\exp(f_n g_n)} &= \frac{\exp\left(g_n \log(1+f_n)\right)}{\exp\left(f_n g_n\right)} \\ &= \exp\left(f_n^2 g_n \sum_{l=0}^{\infty} \frac{1}{l+2}(-1)^{l+1} f_n^l \right) \end{aligned}$$

for all sufficiently large n. We have shown that

$$\lim_{n\to\infty} \sum_{l=0}^{\infty} \frac{1}{l+2}(-1)^{l+1} f_n^l = -\frac{1}{2}.$$

Since $f_n^2 g_n$ approaches zero as n goes to infinity, by using the continuity of $\exp(\cdot)$, we have

$$\begin{aligned} \lim_{n\to\infty} \frac{(1+f_n)^{g_n}}{\exp(f_n g_n)} &= \exp\left(\lim_{n\to\infty} f_n^2 g_n \lim_{n\to\infty} \sum_{l=0}^{\infty} \frac{1}{k+2}(-1)^{l+1} f_n^l \right) \\ &= 1, \end{aligned}$$

which completes the proof of the result.

For the outage probability, we have

$$\beta_n = (1 - \theta_n(-\epsilon \log\log n))^n.$$

Now, by using Lemma 5.5 and the result proved in this appendix, we have

$$\beta_n = O\left(\mathrm{e}^{-(\log n)^{\epsilon} k_n}\right) = O\left(\frac{1}{(\log n)^{\epsilon}}\right), \tag{5.46}$$

which completes the proof of the lemma.

5.9.8 Proof of Theorem 5.3

To start with, the system will perform its best on the rate dimension if we only let the MUs in $l^{\star}$ to feedback such that they satisfy the feedback constraint λ. That is, we set ∞ as the threshold for all groups except $l^{\star}$. Clearly, this will be an upper bound on the achievable rate. Similarly, we will achieve the worst performance if we set a threshold which satisfies the feedback constraint λ for group $l^{\flat}$, and ∞ as the threshold for all other groups. This will be a lower bound on the achievable rate.

Since $\bar{n} \to \infty$ when $n \to \infty$, $\gamma_1^{\star}(\bar{n}, l^{\star})$ and $\gamma_1^{\star}(\bar{n}, l^{\flat})$ converge in distribution to the Gumbel distribution for the best case and the worst case, respectively. Hence, we have the average number of MUs feeding back fixed at e^{-x} from Lemma 5.4, and the outage probability of a beam at $\exp(-\mathrm{e}^{-x})$ from the Poisson approximation, implying we have $O(1)$ feedback with

$$\lim_{n\to\infty} \Lambda\left(\boldsymbol{\mathcal{T}}^{(n)}\right) = M\mathrm{e}^{-x} \tag{5.47}$$

for both cases as well. From Theorem 5.1, we have the trade off curve

$$C^{\star}(\lambda) = 1 - \exp\left(-\frac{\lambda}{M}\right),$$

for both the worst case and the best case. Therefore, for any intermediate threshold setting which achieves the same $O(1)$ feedback constraint, we will end up having the same

tradeoff curve, which completes the proof.

5.9.9 Proof of Theorem 5.6

To start with, since $O\left((\log \bar{n})^{\epsilon}\right)$ MUs will feed back from each group, we have

$$\Lambda\left(\boldsymbol{\mathcal{T}}^{(n)}\right) = O\left((\log \bar{n})^{\epsilon}\right) = O\left((\log n)^{\epsilon}\right)$$

for the scheduling policies.

It is not hard to see that the throughput will scale like $M \log\log n$ at each mini-slot if the BUFAP is used since $\bar{n} \to \infty$ as $n \to \infty$. Therefore, we have

$$\lim_{n\to\infty} \frac{R_n\left(\boldsymbol{\mathcal{T}}^{(n)}\right)}{M \log\log n} = 1 \tag{5.48}$$

for this scheduling policy as well as ensuring fairness.

To prove the same result for RUAP case, first we assume that all the groups have the same SNR $\rho_{\max}$. Now, from Theorem 5.2, we have

$$\lim_{n\to\infty} \frac{R_n\left(\boldsymbol{\mathcal{T}}^{(n)}\right)}{M \log\log n} = 1. \tag{5.49}$$

Then, we assume that all the groups have the same SNR $\rho_{\min}$. Again from Theorem 5.2, we have

$$\lim_{n\to\infty} \frac{R_n\left(\boldsymbol{\mathcal{T}}^{(n)}\right)}{M \log\log n} = 1, \tag{5.50}$$

which completes the proof.

Chapter 6

Optimal Selective Feedback Policies under Peak Feedback Constraints

Opportunistic beamforming is a well-known communication technique that utilizes partial channel state information (CSI) to obtain multiuser diversity gains in the downlink. We focus on the structure of the optimal homogenous threshold feedback policy that maximizes the ergodic downlink sum-rate for opportunistic beamforming under a peak feedback load constraint, which we model by using a multi-packet reception model for the uplink. The users with positive feedback decisions feed their channel states back to the base station simultaneously. The base station can reconstruct all the feedback packets successfully if and only if the random number of MUs feeding back is less than or equal to λ, which is the peak feedback load constraint. We solve the resulting quasi-convex optimization problem by obtaining a formula for the sum-rate maximizing feedback probability. The result holds for many practical fading channel models. While providing insights on the implications of our results in practical systems, we also illustrate the tradeoff between feedback and rate by obtaining the Pareto optimal boundary between feasible and infeasible feedback-rate pairs. We also analyze whether a homogenous threshold level is always rate-wise optimal even for a set of statistically identical users.

6.1 Introduction

OPPORTUNISTIC beamforming is an adaptive signaling technique to reduce the amount of feedback load for wireless vector broadcast channels [48, 97]. In this technique, we opportunistically schedule randomly formed information carrying beams among the mobile users (MUs) by taking partial CSI into account. An important property of opportunistic beamforming is that it achieves the full CSI sum-rate capacity at the downlink to a first order [79]. In this chapter, motivated by such opportunistic communication techniques, we focus on the downlink sum-rate maximization for vector broadcast

channels under *peak* finite feedback constraints on the uplink.

Although opportunistic beamforming reduces the feedback load considerably in comparison to having full CSI, it still requires all the MUs to feed back in its plain implementations. This leads to an impractical linear growth in the feedback load with the number of MUs. There is also a significant waste of communication resources created by the feedback packets from the MUs having no realistic chance of being scheduled for downlink transmission. A potential solution alleviating this impracticality is the use of a selective feedback technique in which only the MUs having good instantaneous channel states are multiplexed on the uplink feedback channel [21,29,37,73,75,76,79].

The results obtained in Chapters 3, 4 and 5 are promising to obtain an $O(1)$ feedback load, on the average, through such a selective feedback mechanism. In particular, we established the structure of selective feedback policies maximizing the vector broadcast sum-rate such that the average number of MUs feeding back is less than a given finite feedback load constraint. However, we assumed an ideal medium-access-control (MAC) layer for contention resolution on the uplink feedback channel, and did not address the likely packet collisions due to a potentially large number of MUs attempting to send their feedback packets back to the base-station.

In this chapter, we consider a multi-packet reception model to resolve collisions from MUs. MUs with positive feedback decisions feed their channel states back to the base station simultaneously. The base station can reconstruct *all* the feedback packets successfully if and only if the random number of MUs feeding back is less than or equal to λ, which is the maximum number of packets that the base station can decode concurrently. The base station utilizes information contained in the decoded feedback packets to assign beams to MUs. We say that a *collision* occurs if the random number of MUs feeding back is greater than λ. In this case, all packets are destroyed together. We will derive the structure of the optimal homogenous threshold feedback policy that maximizes the ergodic downlink sum-rate under the peak feedback load constraint λ imposed by this channel model.

The studied model is general in a sense that $\lambda = 1$ gives us the slotted Aloha collision model [72], and $\lambda > 1$ gives us a special class of a multi-packet reception collision

model called T-out-of-N channels [31, 50]. Such channels can be implemented by using T-out-of-N codes [52]. However, unlike most such channel models, this work does not consider back-off and retransmission issues. Hence, a collision will lead to zero rate. This assumption helps to eliminate the base station from receiving expired information on the instantaneous channel states. Interested readers are referred to [13] and the references therein for more recent work on multi-packet reception. Among them, [107–109] include extensions of multi-packet reception to MIMO wireless local area networks (LAN).

Our contributions and the organization of the chapter are as follows. We explain the multiple access technique for feedback, and formulate the optimization problem in Section 6.2. The base station uses the time-division-multiple-access (TDMA) technique to mediate the process of feedback acquisition. Let M be the number of randomly formed information carrying beams. Then, M mini-slots are allocated for feedback acquisition. All MUs having a positive feedback decision on beam k transmit their feedback packets in the kth mini-slot simultaneously. We call this multiple access model the *Feedback Resource Partitioned Model*. Using the model, we set up the problem of finding the optimal homogenous threshold value which maximizes the ergodic downlink sum-rate, such that the peak feedback load (the number of MUs feeding back in each mini-slot) is less than a constant λ.

Then, in Section 6.3, we show that this is a quasi-convex optimization problem. The quasi-concavity definition cannot be applied directly due to the complexity of the rate expression. Hence, we resort to a first order analysis to prove the quasi-concavity, and to solve the problem and obtain a formula for the optimal feedback probability. The proof of the quasi concavity is distribution independent, *i.e.*, the proof does not depend on the particular statistical model of the wireless channel as long as the resulting SINR distribution is continuous, and the formula for the optimal feedback probability holds for most practical fading distributions such as Rayleigh, Ricean and Nakagami.

In Section 6.4, we make a subtle modification to the multiple access technique given in Section 6.2. This is done by eliminating the TDMA process among the beams in the feedback phase. In this model, the feedback resource is shared between beams, and all the MUs having a positive feedback decision will feedback simultaneously regardless of

the beam. We call this the *Feedback Resource Sharing Model*. We derive the rate expression for this model as well.

Then, in Section 6.5, we apply our results to a Rayleigh fading channel model to provide further insights, and discuss the implications of the derived formulas in practical systems. We provide numerical evidence to the quasi-concavity of rate, and demonstrate the amount of feedback reduction that can be achieved without any noticeable performance degradation in rate by setting the threshold levels optimally. We also illustrate the tradeoff between feedback and rate by obtaining the Pareto optimal boundary between feasible and infeasible feedback-rate pairs. Any feedback-rate pair strictly above the boundary is an infeasible operating point, and a pair strictly below this boundary is a feasible but suboptimal operating point.

In the above analysis, we only focus on the class of homogenous threshold feedback policies when solving the optimal threshold selection problem. However, a question arises of whether a homogenous threshold is always rate-wise optimal. Similar to the justifications done in Chapter 4, a homogenous threshold level seems optimal intuitively because the MUs are statistically identical. However, in Section 6.6, we provide a simple counter example which shows that this intuition is not always true, and for some channel conditions, a better sum-rate can be achieved by setting threshold levels unequally among the MUs. We give reasons for this sub-optimality, and Section 6.7 concludes the chapter.

6.2 System Model and Problem Setup

6.2.1 System Model

In this chapter, we again study the vector broadcast channel model given in Subsection 2.2.1. To recall, the base station has N_t transmit antennas, and each MU is equipped with a single receive antenna. The base station communicates with n MUs through M different beams along the directions of M orthonormal beamforming vectors $\left\{\boldsymbol{b}_k = (b_{1,k}, \ldots, b_{N_t,k})^\top\right\}_{k=1}^{M}$ simultaneously. The beams are assumed to be statistically identical, and MUs experience

i.i.d. channel conditions. $\gamma_{i,m}$ is the SINR at beam m at MU i, and it is given by

$$\gamma_{m,i} = \frac{|\boldsymbol{h}_i^\top \boldsymbol{b}_m|^2}{\rho^{-1} + \sum_{k=1,k\neq m}^{M} |\boldsymbol{h}_i^\top \boldsymbol{b}_k|^2}, \tag{6.1}$$

where ρ is the transmit power per beam, $\boldsymbol{h}_i$ is the vector channel gain between the base station and MU i, Z_i is the unit power (complex) Gaussian background noise. With these normalized parameter selections, ρ also signifies the SNR per beam.

Let $\boldsymbol{\gamma}_i = (\gamma_{1,i}, \ldots, \gamma_{M,i})^\top \in \mathbb{R}_+^M$ represent the SINR vector at MU i. Since the channel gains are i.i.d., the elements of $\boldsymbol{\gamma}_i$ are identically distributed for all $i \in \mathcal{N}$ with a common marginal distribution F, where $\mathcal{N} = \{1, \ldots, n\}$. We will assume that F is continuous, and has the density f with support $\mathbb{R}_+$. Let $\mathcal{M} = \{1, \ldots, M\}$.

6.2.2 Multiple Access Technique for Feedback

We focus on the sum-rate maximization under peak finite feedback constraints, where only a subset of MUs feed back according to a predefined selective feedback policy. To obtain partial CSI from this subset, the base station first broadcasts a feedback request packet (FRP). On retrieval of this packet, the MUs make a feedback decision using a homogenous threshold feedback policy, which we define as follows.

Definition 6.1. *We say $\boldsymbol{\mathcal{T}}$ is a homogenous threshold feedback policy if $\boldsymbol{\mathcal{T}}(\gamma_{k,i})$ generates a feedback packet (FP) containing the* SINR *value $\gamma_{k,i}$ if and only if $\gamma_{k,i} \geq \tau$ for all $k \in \mathcal{M}$ and $i \in \mathcal{N}$. τ is the homogenous threshold value.*

The base station uses the time-division-multiple-access (TDMA) technique to mediate the process of feedback acquisition. More precisely, M mini-slots with length t are allocated for FPs. All MUs having a positive feedback decision on beam k transmit their FPs in the kth mini-slot simultaneously. The values for τ and t can be communicated to the MUs at system initialization. The feedback acquisition process is pictorially represented in Fig. 6.1. We call this multiple access model *Feedback Resource Partitioned Model*, or just partitioned model in short.

It is important to note that a MU having an SINR value above the threshold on two

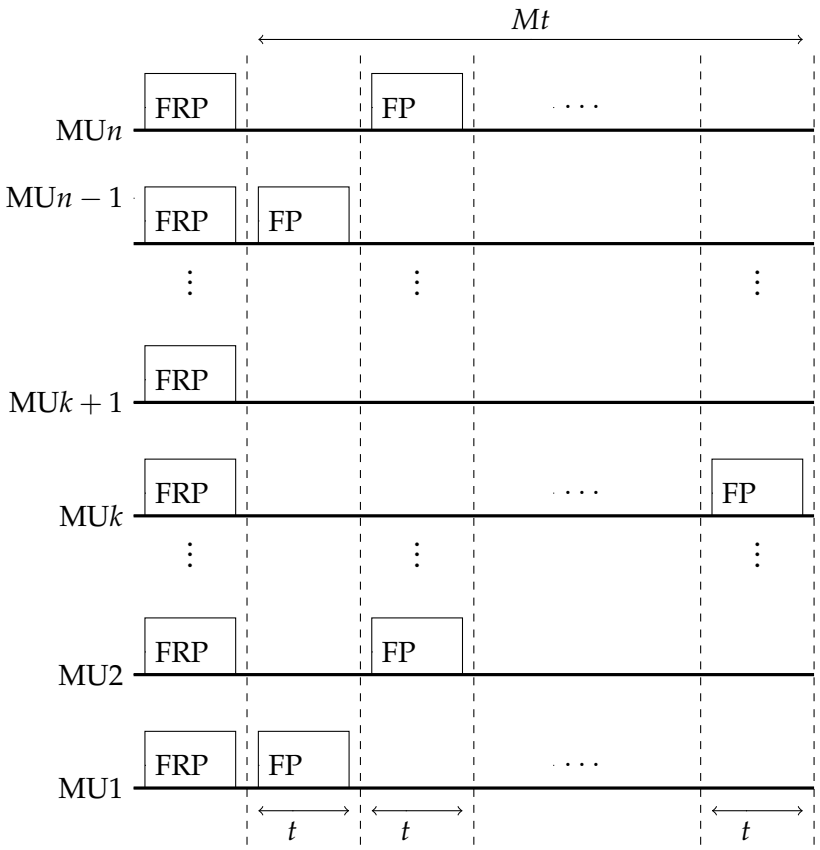


Figure 6.1: The Multiple Access Technique for Feedback: The Feedback Resource Partitioned Model.

or more beams will end up sending more than one FP according our definition of the feedback policy. This can be considered as a waste of communication resources since the MUs can piggy-back the additional data with the first FP. However, it has been shown and discussed in detail in Chapter 3 (Lemma 3.5) that this will not happen if the threshold value is above one (*i.e.*, 0 [dB]), which holds for almost all practical communication scenarios.

6.2.3 The Optimization Problem

We define the truncated SINR on beam m at MU i as

$$\bar{\gamma}_{m,i} = \gamma_{m,i} 1_{\{\gamma_{m,i} \geq \tau\}},$$

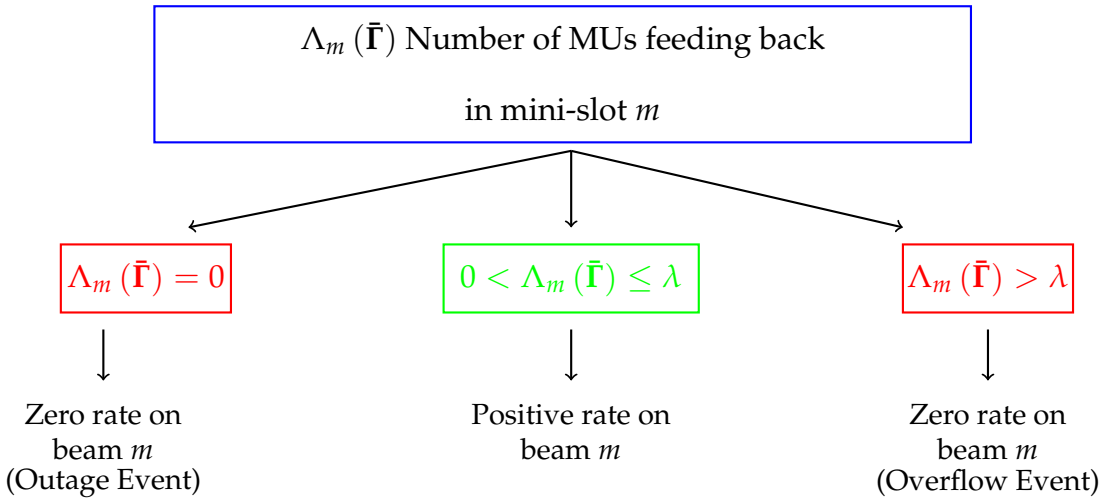


Figure 6.2: The channel model.

and

$$\bar{\mathbf{\Gamma}} = (\bar{\gamma}_{m,i})_{m\in\mathcal{M},i\in\mathcal{N}}$$

is the system-wide M-by-n truncated SINR matrix containing the truncated SINR values of all MUs in the system. Given a threshold value τ, $\Lambda_m(\bar{\mathbf{\Gamma}})$ denotes the random number of MUs feeding back on beam m, or alternatively, the random number of FPs transmitted in mini-slot m. We have

$$\Lambda_m(\bar{\mathbf{\Gamma}}) = \sum_{i=1}^{n} 1_{\{\gamma_{m,i}\geq\tau\}},$$

and it is our performance measure along the feedback dimension.

The base station will decode the FPs with probability one to retrieve the SINR information if and only if the number of FPs transmitted in the respective mini-slot is less than the feedback constraint λ. On the other hand, a collision will occur and all packets will be lost with probability one if the number of FPs exceed λ. There will be zero rate on beam m if the base station has no SINR information regarding the respective beam. This happens on the *overflow event*, where more than λ MUs feed back, and also on the *outage event*, where no MU feeds back. These ideas are pictorially represented in Fig. 6.2.

The base station selects the MU with the highest SINR on each beam for downlink data transmission to maximize the instantaneous communication rate. Then, the downlink ergodic sum-rate achieved for a threshold value τ and a feedback limit λ is given

by

$$\begin{aligned} R_\lambda(\tau) &= \mathsf{E}\left[r_\lambda(\tau, \bar{\mathbf{\Gamma}})\right] \\ &= \mathsf{E}\left[\sum_{m=1}^{M} \log\left(1 + \max_{1\le i\le n} \bar{\gamma}_{m,i}\right) 1_{\{\Lambda_m(\bar{\mathbf{\Gamma}})\le\lambda\}}\right], \end{aligned} \tag{6.2}$$

where $r_\lambda(\tau, \bar{\mathbf{\Gamma}})$ is the *instantaneous* sum-rate achieved for a given τ and λ, and the expectation is taken over the random truncated SINR matrices. Note that the rate on beam m is automatically equal to zero due to truncation if $\Lambda_m(\bar{\mathbf{\Gamma}}) = 0$[1]. $R_\lambda^m(\tau)$ and $r_\lambda^m(\tau, \bar{\mathbf{\Gamma}})$ denote the ergodic rate and the instantaneous rate on beam m, respectively. Also, the ergodic sum-rate achieved on an event $\mathcal{A}$ is written as

$$R_\lambda(\tau, \mathcal{A}) = \mathsf{E}\left[r_\lambda(\tau, \bar{\mathbf{\Gamma}})\, \mathbf{1}_{\mathcal{A}}\right].$$

We will use $R_\lambda(\tau)$ as the performance measure along the rate dimension.

Our main optimization problem is to determine the optimal homogenous threshold value that maximizes the downlink sum-rate such that the number of MUs feeding back on each beam at each channel realization is less than or equal to λ. Since the rate achieved on the outage and overflow events is zero according to (6.2), we can formulate this problem as an unconstrained optimization problem. We can also view the sum-rate as a function of the feedback probability $p = \mathsf{Pr}\{\gamma_{i,m} \ge \tau\}$ without loss of generality since the probability density function of the SINR is already assumed to have $\mathbb{R}_+$ as its support. Therefore, there is a one-to-one correspondence between the threshold value τ and the feedback probability p, *i.e.*, $\tau = F^{-1}(1-p)$. Using these justifications, the main optimization problem to be solved can be represented as

$$\underset{p\in[0,1]}{\text{maximize}} \quad R_\lambda(p) \ . \tag{6.3}$$

If $\lambda = n$, $R_\lambda(p)$ strictly increases as a function of p, $0 \le p \le 1$, and attains its max-

[1]Note that the base station does not have access to any CSI on the feedback outage or overflow events. Without any CSI, reliable communication is still possible if we can average over very large time-scales for all MUs. The extra rate term to be added to (6.2) in this case would not affect our analysis in remainder of the chapter, and therefore is omitted for simplicity.

imum at $p = 1$ since the probability of the overflow event is equal to zero in this case. The system makes use of available communication resources in the best possible way by setting the feedback probability of all MUs to one, *i.e.*, by minimizing the outage event probability. However, when $\lambda < n$, $R_\lambda(0)$ and $R_\lambda(1)$ are both equal to zero since the outage and overflow events will happen with probability one for $p = 0$ and $p = 1$, respectively.

When p is increased from zero to one for $\lambda < n$, $R_\lambda(p)$ firstly increases with p since increasing p reduces the probability of the outage event. On the other hand, increasing p increases the probability of the overflow event in this case. Therefore, the loss created by the overflow event will start to dominate the gain obtained by reducing the outage event probability for large values of p, which eventually causes $R_\lambda(p)$ to decrease and tail off after some certain feedback probability. Therefore, intuitively, $R_\lambda(p)$ will first increase with p up to a certain probability, and then decrease, making $R_\lambda(p)$ quasi-concave over p for $\lambda < n$. In the next section, we will make these intuitive quasi-concavity ideas formal, and solve the resulting optimization problem by obtaining an expression for the optimal feedback probability $p^\star$, which maximizes the downlink ergodic sum-rate for a given feedback constraint λ.

6.3 Selecting The optimal Feedback probability

To solve the optimization problem given in (6.3), it is enough to focus only on the first beam since the beams are statistically identical and $R_\lambda(p)$ can be written as

$$R_\lambda(p) = M\mathsf{E}\left[\log\left(1 + \max_{1\le i\le n} \bar{\gamma}_{1,i}\right) 1_{\{\Lambda_1(\bar{\mathbf{\Gamma}})\le\lambda\}}\right]. \tag{6.4}$$

By using this idea, we will further evaluate the downlink sum-rate through the following lemma.

Lemma 6.1. *The ergodic sum-rate of the system in consideration is given by*

$$R_\lambda(p) = M\sum_{k=1}^{\lambda}\binom{n}{k}k(1-p)^{n-k}I_{k-1}(p), \tag{6.5}$$

where

$$I_{k-1}(p) = \int_{F^{-1}(1-p)}^{\infty} \log(1+x)\left[F(x)-(1-p)\right]^{k-1} dF(x). \tag{6.6}$$

Proof. See Appendix 6.8.1. □

The proof of the quasi-concavity of rate over p requires some more effort since the quasi-concavity definition cannot be applied directly. Hence, we will analyze the first derivative of $R_\lambda(p)$ as a function of p, which is formally obtained in the following lemma.

Lemma 6.2. *The first derivative of the sum-rate $R_\lambda(p)$ is given by*

$$R'_\lambda(p) = M\left[n(1-p)^{n-1}\log\left(1+F^{-1}(1-p)\right) - g_\lambda(p)\right], \tag{6.7}$$

where

$$g_k(p) = \binom{n}{k} k(n-k)(1-p)^{n-k-1} I_{k-1}(p) \tag{6.8}$$

for $k \in \{1,\ldots,\lambda\}$.

Proof. See Appendix 6.8.2. □

By manipulating the expression for the first derivative of the sum-rate, we will prove that the ergodic sum-rate is a quasi-concave function of p. Then, by using this behavior, we will obtain an expression for the optimal feedback probability $p^\star$, which is formally stated in the following theorem. We focus only on the case where $\lambda < n$. For $\lambda = n$, $p^\star$ is equal to 1 as discussed above.

Theorem 6.1. *Let $\lambda < n$. Then, the optimal feedback probability per beam $p^\star$ for the system in consideration is given by the solution to*

$$cI_{\lambda-1}(p^\star) - (1-p^\star)^\lambda \log\left(1+F^{-1}(1-p^\star)\right) = 0, \tag{6.9}$$

where $c = \lambda\binom{n-1}{\lambda}$.

Proof. The stationary points of $R_\lambda(p)$ for $p \in (0,1)$ can be obtained from (6.7), and they satisfy

$$Mn(1-p)^{n-1}\left[\log\left(1+F^{-1}(1-p)\right)-\frac{g_\lambda(p)}{n(1-p)^{n-1}}\right]=0.$$

After substituting for $g_\lambda(p)$ from (6.8), any p which satisfies

$$(1-p)^\lambda \log\left(1+F^{-1}(1-p)\right)=cI_{\lambda-1}(p) \tag{6.10}$$

will be a stationary point since $Mn(1-p)^{n-1}>0$ for all $p\in(0,1)$. Let

$$G_1(p)=cI_{\lambda-1}(p),$$

and

$$G_2(p)=(1-p)^\lambda \log\left(1+F^{-1}(1-p)\right).$$

We have $G_1(0)=0$, and

$$G_1(1)=c\int_0^\infty \log(1+x)\,[F(x)]^{\lambda-1}\,dF(x).$$

Note that $G_1(1)>0$. We also have

$$G_1'(p)=\begin{cases} c(\lambda-1)\,I_{\lambda-2}(p) & \text{if } \lambda\geq 2 \\ c\log\left(1+F^{-1}(1-p)\right) & \text{if } \lambda=1 \end{cases}, \tag{6.11}$$

which is again strictly positive for any $p\in(0,1)$. Hence, $G_1(p)$ is strictly increasing function of p, with end points $G_1(0)=0$ and $G_1(1)>0$. Similarly, we have $G_2(1)=0$ and $\lim_{p\to 0} G_2(p)=\infty$. Its first derivative is equal to

$$G_2'(p)=-(1-p)^{\lambda-1}\left[\frac{(1-p)}{(1+F^{-1}(1-p))\,f\left(F^{-1}(1-p)\right)}+\lambda\log\left(1+F^{-1}(1-p)\right)\right],$$

which is strictly negative for any $p\in(0,1)$. Therefore, $G_2(p)$ is strictly decreasing. The

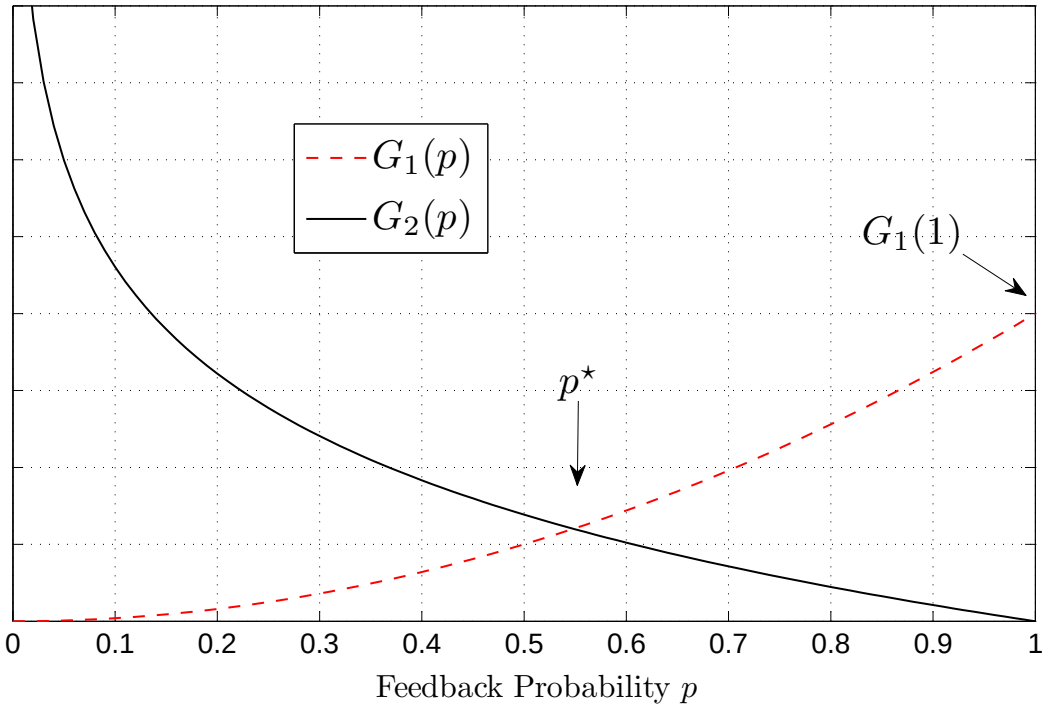


Figure 6.3: A plot illustrating the behavior of $G_1(p)$ and $G_2(p)$.

behavior of $G_1(p)$ and $G_2(p)$ is illustrated in Fig. 6.3.[2] Hence, $G_1(p)$ and $G_2(p)$ intersect at a unique value of p, which implies that there is only one stationary point of $R_\lambda(p)$ for $p \in (0,1)$. Since $\lim_{p\to 0} R'_\lambda(p) > 0$, $R_\lambda(0) = 0$ and $R_\lambda(1) = 0$, this stationary point corresponds to the global maximum of $R'_\lambda(p)$. These arguments also indicate that the sum-rate strictly increases up a point $p^\star$, and then, strictly decreases. Therefore, the sum-rate is a quasi-concave function of p, which completes the proof. □

It is important to note that the proven quasi-concavity property of the sum-rate is distribution independent, and the result in Theorem 6.1 can be used to find the optimal feedback probability for most practical fading distributions such as Rayleigh, Ricean and Nakagami. However, substituting for F in (6.9) using a particular fading distribution may not necessarily lead to a closed form expression for $p^\star$. In such cases, most of the common root finding algorithms can be easily used to find the unique zero crossing since (6.9) is a strictly increasing function of p. This will still be much efficient computationally, compared to doing a line search over p on (6.5). In the next section, we will analyze a

[2] The plot is just given to conceptualize the behavior of the two functions.

slightly different model, where we do an alteration to the multiple access technique for feedback.

6.4 Feedback Resource Sharing Model

A possible modification of the problem setup presented in this chapter can be achieved by eliminating the TDMA process among the beams in the feedback phase. In this model, the feedback resource is shared between beams: we would have a single time slot with increased peak feedback capacity. Thus, we call it the *Feedback Resource Sharing Model*, or just sharing model in short. This means, we have a limit on the total number of MUs feeding back on all the beams, *i.e.*, we consider

$$\sum_{m\in\mathcal{M}} \Lambda_m(\bar{\mathbf{\Gamma}}) \leq \lambda_{sm},$$

where λ_{sm} is the peak feedback constraint of the sharing model (recall, the peak feedback constraint of the partitioned model is λ). The operational difference of this multiple access technique to the one proposed in Subsection 6.2.2 is illustrated in Fig. 6.4. Now all the MUs will feed back their FPs simultaneously regardless of the beam.

Although the multiple access technique looks simpler compared to the partitioned case, the analysis of this model is much harder due to the complexity of the expression of rate.We formally obtain this expression through the following lemma.

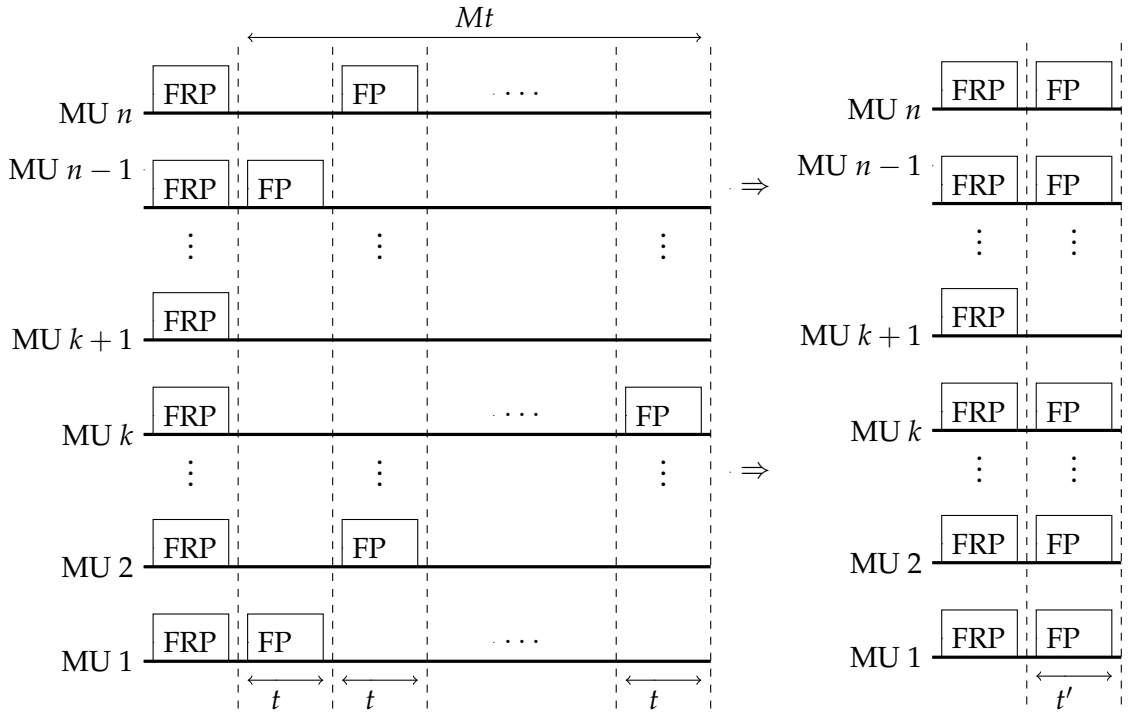
Lemma 6.3. *The ergodic sum rate for the channel sharing model is given by*

$$R_{\lambda_{sm}}(p) = M \sum_{j_1=1}^{\min(\lambda_{sm},n)} \sum_{j_2=0}^{\min(\lambda_{sm},n)-j_1} \cdots \sum_{j_M=0}^{\min(\lambda_{sm},n)-\sum_{k=1}^{M-1} j_k} \prod_{m=1}^{M} \binom{n-\sum_{k=1}^{m-1} j_k}{j_m} \times j_1 p^{\sum_{k=2}^{M} j_k} (1-p)^{Mn-\sum_{k=0}^{M-k} j_{k+1}} I_{j_1-1}(p), \quad (6.12)$$

where $\tau > 1$, and I is defined as in (6.6).

Proof. See Appendix 6.8.3. □

Again, intuitively, the rate is quasi concave because all the arguments we made under

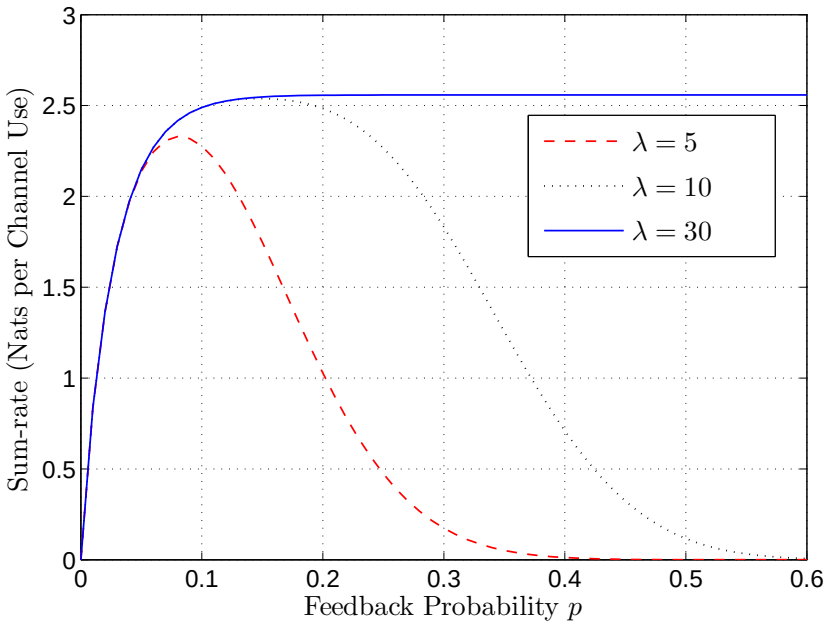


(a) Feedback Resource Partitioned Model.

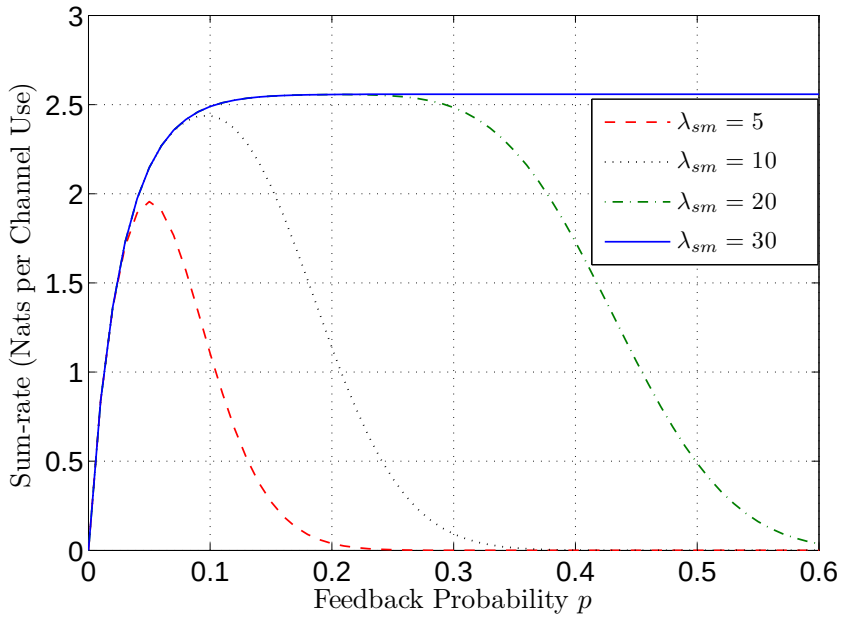
(b) Feedback Resource Sharing Model.

Figure 6.4: The comparison of the operation in the MAC layer for a given channel realization.

the partitioned case to justify this behavior still hold for the sharing model. The quasi-concavity definition cannot be applied directly for the proof of the quasi concavity of this model as well. However, unlike the partitioned case, where we analyzed the first derivative for the proof of quasi concavity, the first derivative for this case turns out to be complex and mathematically untractable. This makes the proof of the quasi concavity of this model extremely hard. Therefore, we resort to numerical evaluations to back our intuition, which will be presented in the next section. On top of this, due to the complexity of the first derivative, doing a line search over p on (6.12) might be more beneficial computationally to find the optimal feedback probability, compared to finding the unique zero crossing using the first derivative as we did in the partitioned case. In the next section, we will provide some numerical evaluations and insights into our results using a Rayleigh fading channel model.



(a) Feedback Resource Partitioned Model



(b) Feedback Resource Sharing Model

Figure 6.5: A plot illustrating the behavior of $R_{\lambda}(\tau)$ and $R_{\lambda_{sm}}(\tau)$ for different feedback resources, where $n = 30$, $M = 2$ and $\rho = 1$.

6.5 Discussion of Results: Rayleigh Fading Channels

For this numerical study, we will use Rayleigh fading channels, which is one of the most commonly used channel models in the literature, *e.g.*, see [9, 82, 83], and closely approximates measured data rates in densely populated urban areas [14]. We assume transmitted signals of unit power, and Rayleigh distributed channel fading coefficients of unit power *i.e.*, $h_{k,i}, k = 1, \ldots, N_t$ and $i = 1, \ldots, n$, are assumed to be i.i.d. with the common distribution $\mathcal{CN}(0,1)$, where $\mathcal{CN}(\mu, \sigma^2)$ represents the *circularly-symmetric complex Gaussian* distribution with mean μ and variance σ^2. Recall that the background noise is the unit power (complex) Gaussian noise, and therefore ρ is interpreted as the average SNR.

For this channel model, the SINR distribution function F and the associated probability density function f can be given as

$$F(x) = 1 - \frac{e^{-\frac{x}{\rho}}}{(x+1)^{M-1}} \tag{6.13}$$

and

$$f(x) = \frac{e^{-\frac{x}{\rho}}}{(x+1)^{M}} \left[\frac{1}{\rho}(x+1) + M - 1\right], \tag{6.14}$$

respectively [79]. F^{-1} for this model is given by

$$F^{-1}(x) = \begin{cases} -1 + (M-1)\rho W\left(\frac{\exp\left(\frac{1}{(M-1)\rho}\right)}{(M-1)\rho}(1-x)^{\frac{1}{1-M}}\right) & \text{if } M \geq 2 \\ -\rho \log(1-x) & \text{if } M = 1 \end{cases},$$

where $x \in [0,1]$ and W is the Lambert W function given by the defining equation

$$W(x)\exp(W(x)) = x$$

for $x \geq -\frac{1}{e}$ [74].

Firstly, we study the behavior of the sum-rate for both the sharing and partitioned cases. As shown in Fig. 6.5, the sum-rate is quasi-concave over p for both the cases, and is strictly increasing when $\lambda = n$, which is in line with our arguments above. When

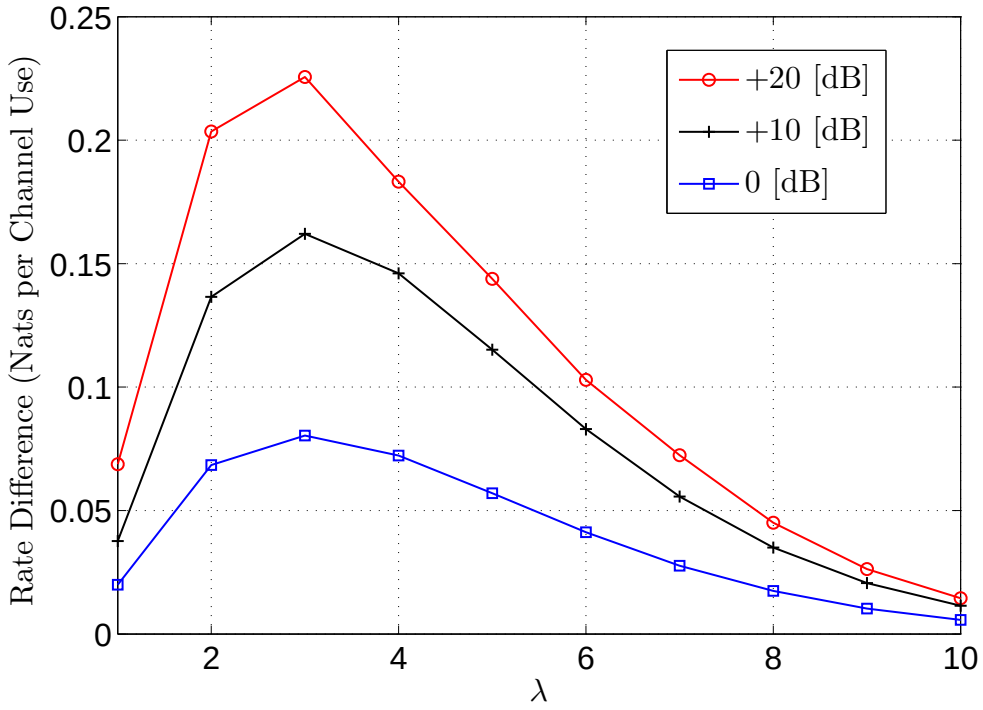


Figure 6.6: A plot illustrating the difference in sum-rate of the two models when $\lambda_{sm} = M \times \lambda$, where $M = 2$ and $n = 20$.

we compare the two cases, we can see that the partitioned case does better if $\lambda_{sm} = \lambda$. This is obviously because more MUs feed back when a partitioned model is used with this resource allocation. Therefore, for a more fair comparison in downlink sum-rate, we consider the case $\lambda_{sm} = M \times \lambda$, where $M = 2$ for this particular figure. We can see that now, the sharing case is superior in terms of sum-rate. However, this is at the expense of higher feedback resources in the uplink. It is also interesting to note that the difference in sum-rate of the two models is considerable only when the feedback resources are limited. A clearer illustration of this difference is given in Fig. 6.6, where we have numerically evaluated the difference in sum-rates for different average SNR values. For the sake of fairness in this comparison, we have set $\lambda_{sm} = M \times \lambda$. The difference firstly increases with λ, and then decreases when the feedback resources are further increased. When the feedback resources are high, the selection of the multiple access technique has only a small effect on the sum-rate. This brings us to the realization that the resource allocation matters the most when the resources are limited.

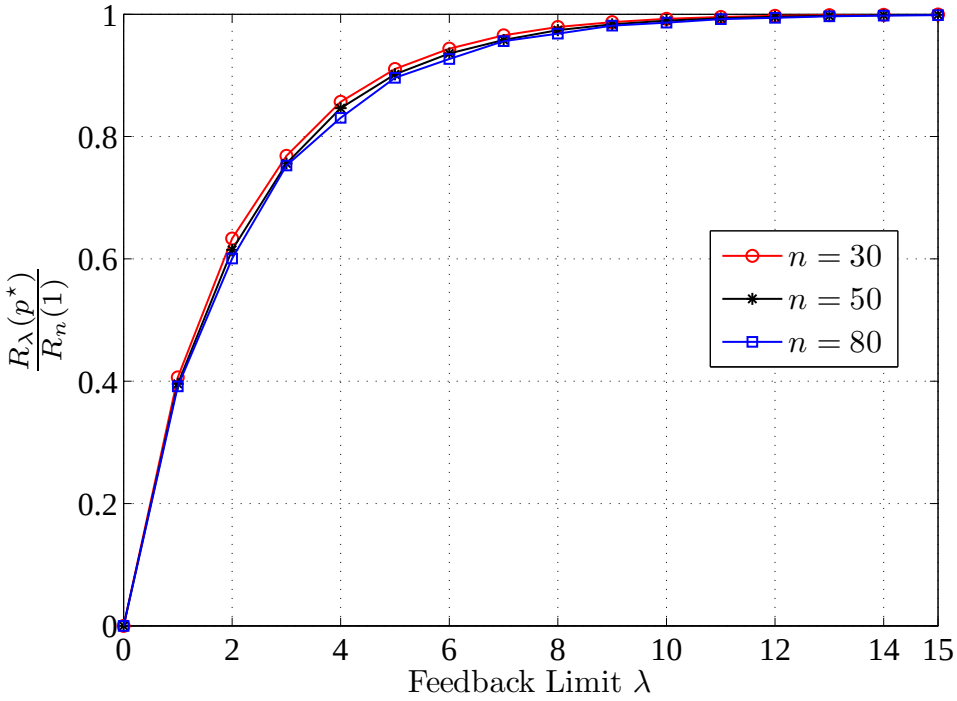


Figure 6.7: A plot illustrating the tradeoff between sum-rate and feedback for different values of n, where $M = 2$ and $\rho = 1$.

Now that we have compared the two models, we will focus more on the partitioned model in the remaining parts of the chapter. We can do similar analysis and discussions for the sharing model, which we will skip to avoid repetition. Coming back to Fig. 6.5(a), we can observe that the difference between the maximum sum-rates for $\lambda = 10$ and $\lambda = n = 30$ is very small. This observation is more clearly depicted in Fig. 6.7, where we have plotted the ratio of sum-rates achieved with $(R_\lambda(p^\star))$ and without $(R_n(1))$ thresholding as a function of λ. The plot, in fact, represents the feedback-rate tradeoff for opportunistic beamforming under peak feedback load constraints for different MU levels, and also represents the amount of feedback reduction that can be achieved by setting the threshold levels according to Theorem 6.1 without any noticeable performance degradation in rate.

The outage and overflow event probabilities are strictly positive when thresholds are optimally set to meet the feedback constraint λ, $\lambda < n$, which causes a rate loss. However, on inspecting Fig. 6.7, we see that there is almost no rate loss if λ is greater than 12, irrespective of the user level. This observation can be explained as follows. The base station

communicates with only the best MU on each beam. Therefore, a perfect feedback policy in terms of optimal usage of the uplink capacity is a policy which allows only the best MU on each beam to feed back. However, since there is no coordination among the MUs, this is impossible to achieve in a distributed manner, and setting $\lambda = 1$ results in a large degradation in sum-rate as shown in the figure, mainly because such a small constraint on the feedback load increases the probability of the outage and overflow events. Therefore, the feedback constraint should be relaxed (or, the multipacket reception capability should be improved) in a way which ensures the best MU feeds back without causing an overflow event. The tails of the distribution of the random number of MUs requesting each beam decays to zero exponentially fast. This means that there is a high probability of the best MU feeding back without causing an overflow when λ is above a certain value, which is 12 in this case. Increasing n increases the probability of the overflow event. This results in the downwards shift of the curves in Fig. 6.7 when n is increased, although it is negligibly small.

Some further comments on Fig. 6.7 are in order. The curves in this figure represent the Pareto optimal boundary between feasible and infeasible feedback-rate pairs, which can be achieved by setting a threshold level in accordance with Theorem 6.1. Any feedback-rate pair above the boundary is an infeasible operating point, which means that the represented sum-rate cannot be achieved subject to the feedback restriction. If a pair is below this boundary, then it is a feasible operating point, which means that the sum-rate can be achieved without violating the feedback constraint. However, a point strictly below the boundary is suboptimal in the sense that the same sum-rate can be achieved with strictly less feedback or a better sum-rate can be achieved while maintaining the same feedback level.

6.6 Optimality of a Homogenous Threshold

In the above analysis, we have only focussed on the class of homogenous threshold feedback policies when studying the optimal threshold selection. However, a question arises of whether a homogenous threshold is always rate-wise optimal. Intuitively, it seems op-

timal because the MUs are statistically identical. However, our next example shows that this is not always true, *i.e.*, sometimes a better sum-rate can be achieved by setting non homogenous threshold levels among the MUs.

We focus on a three-user system, where $M = 1$, and we set λ at 2. Let τ_1, τ_2 and τ_3 be the threshold values, and p_1, p_2 and p_3 be the feedback probabilities of MU 1, 2 and 3, respectively. The rate expression for this system is obtained through the following lemma.

Lemma 6.4. *The downlink ergodic rate of a single beam system with three MUs feeding back using a non-homogenous feedback policy, and having a feedback constraint of* 2 *is given by*

$$R_2\left(p_1,p_2,p_3\right) = \sum_{i=1}^{3}\prod_{k=1,k\neq i}^{3}\left(1-p_k\right)\int_{F^{-1}(1-p_i)}^{\infty}\log\left(1+x\right)dF(x) + \sum_{i=1}^{2}\sum_{j=i+1}^{3}\left(1-p_k\right)_{k\neq i,j}\int_{\max\left(F^{-1}(1-p_i),F^{-1}(1-p_j)\right)}^{\infty}\log\left(1+x\right)dG_{i,j}(x), \quad (6.15)$$

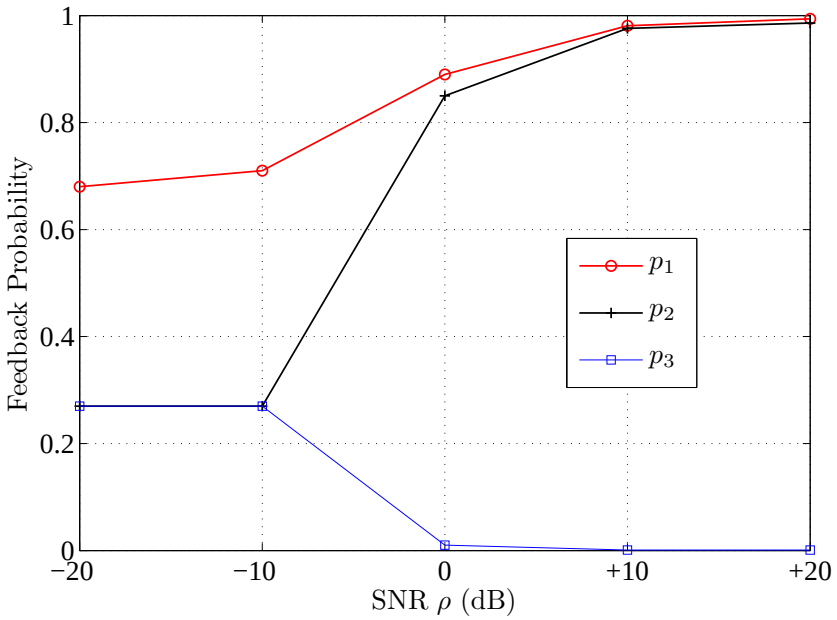
where

$$G_{i,j}(x) = \left[F(x)-F(\tau_i)\right]\left[F(x)-F(\tau_j)\right]. \quad (6.16)$$

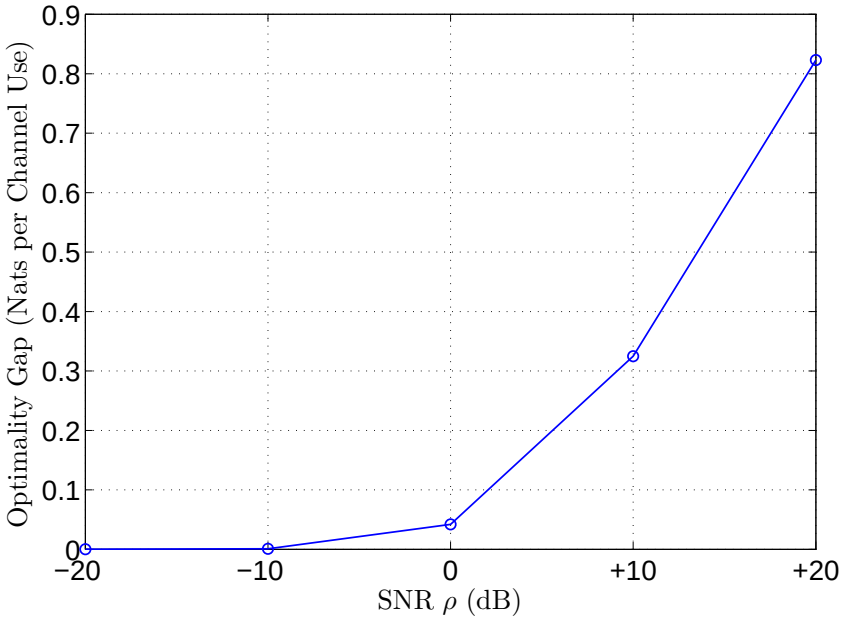
This lemma can be proved by following the lines of the proofs of Lemma 6.1 and Lemma 6.3. Therefore, we skip it to avoid repetition.

We now can do an exhaustive search over p_1, p_2 and p_3 on (6.15) to find the set of optimal feedback probabilities. The results are illustrated in Fig. 6.8(a). As the results show, a homogenous threshold level is clearly sub-optimal for this system. At low SNRs, two of the MUs use the same threshold while the other MU uses a lower threshold value, which leads to a higher feedback probability. As we increase the average SNR, the feedback probabilities of two of the MUs go to one, and the feedback probability of the other goes to zero.

This phenomenon can be explained as follows. Let β_{out} be the outage event probability, and let β_{ovr} be the overflow event probability. The sum-rate in this case can be written as



(a) Optimal feedback probabilities



(b) The optimality gap created by using a homogenous threshold

Figure 6.8: Plots illustrating the sub-optimality of homogenous thresholds for different values of ρ, where $n = 3$ and $\lambda = 2$.

$$R_2(\tau_1, \tau_2, \tau_3) = (1 - \beta_{out} - \beta_{ovr}) \times \mathsf{E}\left[\log\left(1 + \max_{i=1,2,3} \gamma_{1,i} 1_{\{\gamma_{1,i} \geq \tau_i\}}\right) \middle| \text{ No Outage or Overflow } \right]. \quad (6.17)$$

Two key factors affect this rate expression. The first one of them is the maximization operation inside the logarithm, which affects the power gain achieved by the means of multiuser diversity. The gain is likely to be higher if the maximization is done over a larger set of MUs. Therefore, if gaining through multiuser diversity is the objective, the base station would prefer if all the three MUs were feeding back with equal probability. The second factor is the degrees-of-freedom gain represented by the $(1 - \beta_{out} - \beta_{ovr})$ term. It is not hard to see that smaller the β_{out} and β_{ovr}, the higher the degrees-of-freedom gain that we achieve. The choice of thresholds affects both gains, and the interplay between them determines how we set thresholds to maximize the downlink sum-rate.

According to fig. 6.8(a), the degrees of freedom gain dominates the rate equation for the whole range of SNR values considered in this figure. Thus a homogenous threshold is suboptimal. It is also important to notice that the effects of multiuser diversity gains further diminishes when ρ is increased. This can be explained using (6.17) as well. When the SNR is low, the sum-rate increases *almost* linearly with the power gain, due to the logarithm. Therefore, the effect of the multiuser diversity gain is at its maximum at low SNR values. As a result, we have at least two MUs with the same threshold value. In the mean time, the system increases the feedback probability of the other MU to decrease the probability of the outage event.

In the high SNR regime, on the other hand, the power gain can only provide a logarithmic increase in the sum-rate, *i.e.*, the law of diminishing returns. Hence, the system will do better by maximizing the degrees-of-freedom gain, instead of the multiuser diversity gain. We can see that the multiuser diversity gain is almost neglected at high SNR values. The system tries to minimize the probability of the overflow event by sending the feedback probabilities of two of the MUs to one, and the other to zero. The feedback probabilities going to one simultaneously reduce the outage probability as well. The optimality gap, or the rate loss created by using a homogenous threshold instead of setting the thresholds optimally as in Fig. 6.8(a) is shown in Fig. 6.8(b). We can see that the rate

loss increases when the SNR is increased.

To shed more light on the above explanation, we do an extension on the above studied three-user system. We keep λ fixed and increase the number of MUs in the system. Obviously, this increase in the number of MUs increases the ability of the system to gain through multiuser diversity. We show that as we increase the ability of the base station to gain through multiuser diversity, the system resorts to a homogenous threshold level. We consider two groups of MUs. The first group $G1$ has n_1 MUs with a feedback probability of p_1^g and a homogenous threshold value τ_1^g. The second group $G2$ has n_2 MUs with a feedback probability of p_2^g and a homogenous threshold value τ_2^g. The only difference of a MU from $G2$ to a one in $G1$ is the feedback probability. We keep λ fixed at 2, similar to the previous example. We formally obtain an expression for the rate of this system through the following lemma.

Lemma 6.5. *The downlink ergodic rate for the system in consideration is given by*

$$\begin{aligned} R_2\left(p_1^g, p_2^g\right) = &\sum_{i=1}^{2} n_i \left(1-p_i^g\right)^{n_i-1} \left(1-p_k^g\right)_{k\neq i}^{n_k} \int_{F^{-1}(1-p_i^g)}^{\infty} \log\left(1+x\right) dF(x) \\ &+ \sum_{i=1}^{2} \binom{n_i}{2} \left(1-p_i^g\right)^{n_i-2} \left(1-p_k^g\right)_{k\neq i}^{n_k} \int_{F^{-1}(1-p_i^g)}^{\infty} \log\left(1+x\right) dG_{i,i}(x) \\ &+ n_1 n_2 \prod_{k=1}^{2} \left(1-p_k^g\right)^{n_k-1} \int_{\max\left(F^{-1}(1-p_1^g), F^{-1}(1-p_2^g)\right)}^{\infty} \log\left(1+x\right) dG_{1,2}(x), \quad (6.18) \end{aligned}$$

where

$$G_{i,j}(x) = \left[F(x) - F(\tau_i^g)\right] \left[F(x) - F(\tau_j^g)\right]. \quad (6.19)$$

This lemma can be proved by following the lines of the proofs of Lemma 6.1 and Lemma 6.3. Therefore, we skip this proof as well to avoid repetition.

We set $n_1 = 2$. Then, we evaluate (6.18) for different values of n_2, while keeping n_1 constant. The results are illustrated in Fig. 6.9. We start at $n_2 = 1$. This is similar to the three-user scenario studied in this section under Lemma 6.4. The results are similar as well, where the feedback probability of the first group with two MUs is near 0.9, and the feedback probability of the group with 1 MU is almost at zero. However, when we

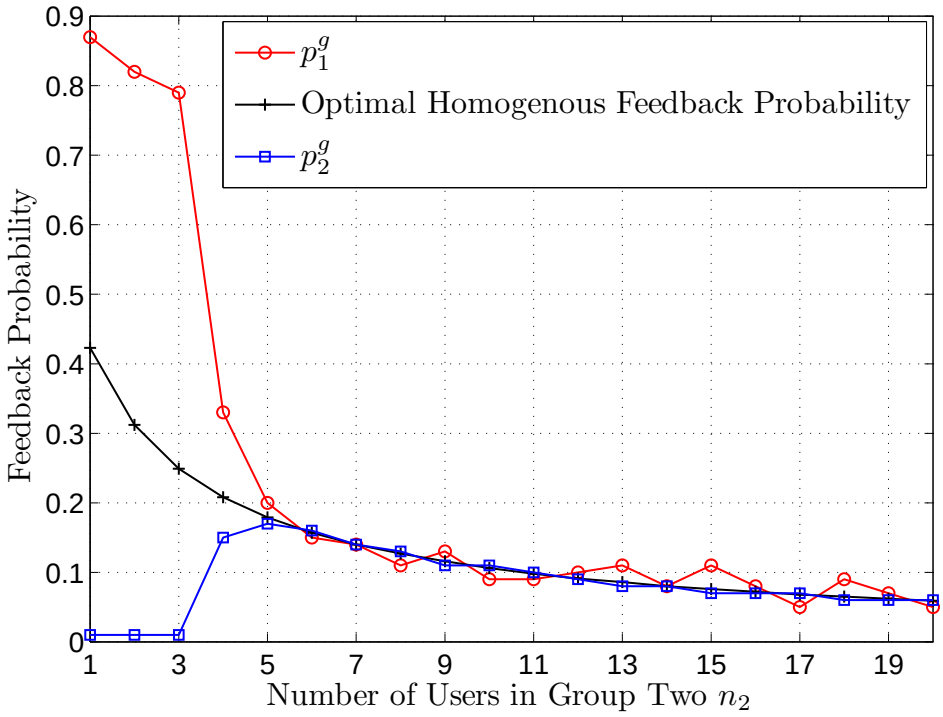


Figure 6.9: A plot illustrating the behavior of p_1^g and p_2^g with n_2, where $n_1 = 2$, $\lambda = 2$, and $\rho = 1$.

increase n_2, we can see p_1^g going down and p_2^g going up. Finally, the probabilities start fluctuating around the optimal homogenous feedback probability obtained using Theorem 6.1, where we set $n = n_1 + n_2$. This means a homogenous threshold level becomes optimal when we increase the number of MUs in the system while keeping λ fixed.

We could give another interpretation to this result as well. When n_2 is increased, the MUs in G_2 starts to dominate the rate because n_1 is fixed at a small value. In other words, the effect of the two MUs in G_1 on the rate diminishes gradually. Therefore, after a certain value of n_2 the system will not gain in rate by allocating a different threshold value to the two MUs in G_1. Therefore, the system starts using a homogenous threshold value.

6.7 Conclusions

In this chapter, we considered a multi-packet reception model on the uplink, where the base station can successfully decode the feedback packets if and only if the number of

feedback packets received is less than a constraint λ. The base station used a TDMA technique among the beams to mediate the process of feedback acquisition. Using this model, we obtained a formula for the downlink sum-rate maximizing feedback probability for any given λ. Then, we made a subtle modification to the multiple access technique for feedback by eliminating the TDMA process among the beams in the feedback phase. We applied our results to the well known Rayleigh fading channel model and discussed the importance of the results. We also obtained the Pareto optimal boundary between feasible and infeasible feedback-rate pairs. Finally, we analyzed the rate-wise optimality of using a homogenous threshold value. We provided a simple counter example which showed that even for a set of statistically identical mobile users, a homogenous threshold level can be sub-optimal for some channel conditions.

6.8 Appendix

6.8.1 Proof of Lemma 6.1

First, consider the event

$$\mathcal{A}_k = \{\bar{\mathbf{\Gamma}} | \Lambda_1 (\bar{\mathbf{\Gamma}}) = k\},$$

which is the event where exactly k MUs feed back on beam one. The probability of this event is

$$p_{\mathcal{A}_k} = \binom{n}{k} p^k (1-p)^{n-k}.$$

Therefore, we can write

$$\begin{aligned} R_\lambda^1 (p, \mathcal{A}_k) &= p_{\mathcal{A}_k} \mathsf{E}\left[r_\lambda^1 (\tau, \bar{\mathbf{\Gamma}}) | \mathcal{A}_k \right] \\ &= p_{\mathcal{A}_k} \int_0^\infty \log(1+x) dH_k(x), \end{aligned} \tag{6.20}$$

where for X_i, $1 \leq i \leq k$, distributed according to F and

$$H_k(x) = \mathsf{Pr}\{\max\{X_1, \ldots, X_k\} \leq x | X_1 \geq \tau, \ldots, X_k \geq \tau\}$$

$$= \left(\frac{F(x) - F(\tau)}{1 - F(\tau)}\right)^k 1_{\{x \geq \tau\}}, \tag{6.21}$$

which is the cumulative distribution function of the maximum of k i.i.d. SINR values given all of them are above τ. Substitution of (6.21) in (6.20) together with $F(\tau) = 1 - p$ gives

$$R_\lambda^1(p, \mathcal{A}_k) = \binom{n}{k} k (1-p)^{n-k} I_{k-1}(p). \tag{6.22}$$

The rate on beam one will be non zero only on events $\mathcal{A}_k$ for $1 \leq k \leq \lambda$. Since these are disjoint events, summing over the average rates achieved on each of these events gives us the ergodic rate on beam one, which completes the proof.

6.8.2 Proof of Lemma 6.2

Using the Leibniz integral rule, we obtain an expression for the first derivative of $R_\lambda^1(p, \mathcal{A}_k)$, which is

$$R_\lambda^1(p, \mathcal{A}_k)' = g_{k-1}(p) - g_k(p)$$

for $k \geq 2$, and

$$R_\lambda^1(p, \mathcal{A}_1)' = \binom{n}{1} (1-p)^{n-1} \log\left(1 + F^{-1}(1-p)\right) - g_1(p)$$

for $k = 1$. Since

$$R_\lambda^1(p) = \sum_{k=1}^{\lambda} R_\lambda^1(p, \mathcal{A}_k),$$

we have

$$R_\lambda'(p) = M \sum_{k=1}^{\lambda} R_\lambda^1(p, \mathcal{A}_k)'.$$

Therefore,

$$\begin{aligned} R_\lambda'(p) &= M\left[R_\lambda^1(p, \mathcal{A}_1)' + \sum_{k=2}^{\lambda} (g_{k-1}(p) - g_k(p))\right] \\ &= M\left[n(1-p)^{n-1} \log\left(1 + F^{-1}(1-p)\right) - g_\lambda(p)\right], \end{aligned}$$

which completes the proof.

6.8.3 Proof of Lemma 6.3

We will consider the rate on beam 1. Take the event

$$\mathcal{A}_{j_1,\ldots,j_M} = \{\bar{\mathbf{\Gamma}} | \Lambda_k(\bar{\mathbf{\Gamma}}) = j_k \ \ \forall k \in \mathcal{M}\},$$

which is the event where exactly j_k MUs feed back on each beam $k \in \mathcal{M}$. Since $\tau > 1$, a MU will not feed back on more than one beam[3]. Therefore the probability of this event can be written as

$$\begin{aligned}\mathsf{Pr}\left\{\mathcal{A}_{j_1,\ldots,j_M}\right\} &= \binom{n}{j_1}\cdots\binom{n-\sum_{k=1}^{M-1} j_k}{j_M} p^{j_1+\ldots+j_M}(1-p)^{n-j_1}\times\ldots\times(1-p)^{n-\sum_{k=1}^{M-1} j_k} \\ &= \prod_{m=1}^{M}\binom{n-\sum_{k=1}^{m-1} j_k}{j_m} p^{\sum_{k=1}^{M} j_k}(1-p)^{Mn-\sum_{k=0}^{M-k} j_{k+1}}.\end{aligned}$$

Therefore, similar to the partitioned case, we can write

$$R^1_{\lambda_{sm}}(p, \mathcal{A}_{j_1,\ldots,j_M}) = \mathsf{Pr}\left\{\mathcal{A}_{j_1,\ldots,j_M}\right\}\int_0^\infty \log(1+x)dH_{j_1}(x), \tag{6.23}$$

where $H_{j_1}(x)$ is defined as in (6.21). Substitution of $H_{j_1}(x)$ in (6.23) together with $F(\tau) = 1-p$ gives

$$R^1_{\lambda_{sm}}(p, \mathcal{A}_{j_1,\ldots,j_M}) = \frac{\mathsf{Pr}\left\{\mathcal{A}_{j_1,\ldots,j_M}\right\}}{p^{j_1}} j_1 I_{j_1-1}(p). \tag{6.24}$$

The rate will be non zero only on events where $\sum_{k=1}^{M} j_k \leq \lambda_{sm}$. Actually, since a MU will not feed back on both the beams, $\sum_{k=1}^{M} j_k \leq \min(\lambda_{sm}, n)$. Therefore we can obtain the rate on beam one by introducing summations over j_k for all $k \in \mathcal{M}$ such that this condition is fulfilled, as follows.

[3]Since the SINRs at a user on different beams are dependent random variables, the event probabilities cannot be written as a product of the feedback probabilities as we did in the partitioned scenario if a MU feeds back on more than one beam.

$$R^1_{\lambda_{sm}}(p) = \sum_{j_1=1}^{\min(\lambda_{sm},n)} \sum_{j_2=0}^{\min(\lambda_{sm},n)-j_1} \cdots \sum_{j_M=0}^{\min(\lambda_{sm},n)-\sum_{k=1}^{M-1} j_k} \prod_{m=1}^{M} \binom{n-\sum_{k=1}^{m-1} j_k}{j_m} \times j_1 p^{\sum_{k=2}^{M} j_k} (1-p)^{Mn-\sum_{k=0}^{M-k} j_{k+1}} I_{j_1-1}(p).$$

This completes the proof since the beams are statistically identical.

Chapter 7
Conclusions and Future Work

7.1 Conclusions

Opportunistic beamforming is an important communication strategy achieving the full CSI sum-rate capacity for vector broadcast channels to a first order by only requiring partial CSI at the base station. Nevertheless, it cannot eliminate the linear growth in the feedback load with increasing numbers of mobile users (MUs) in the network unless a selective feedback policy is implemented for user selection. In this thesis, we focused on a more stringent but practical finite limit on the feedback load, and studied the structure of the sum-rate maximizing decentralized selective feedback policies, and how the resulting sum-rate compare to the sum-rate without any user selection.

First, we set a finite limit on the average number of MUs feeding back. We showed that threshold feedback policies in which MUs compare their beam signal-to-interference-plus-noise-ratios (SINRs) with a threshold for their feedback decisions are always optimal to maximize the downlink sum-rate. This class of policies was studied in many previous works without any formal justification for why they are the right choice for user selection. Our thresholding optimality result provides the formal justification, which holds for all fading channel models with continuous distribution functions.

Having established the optimality of threshold feedback policies, we then studied the optimal threshold selection problem to maximize the sum-rate under the same constraints on feedback. This was a non-convex optimization problem over finite dimensional Euclidean spaces. We solved this problem by identifying an underlying Schur-concave structure in the sum-rate. Specifically, we have obtained sufficient conditions

ensuring the Schur-concavity of the sum-rate, and therefore the rate optimality of homogenous threshold feedback policies. These sufficient conditions were provided for general fading channel models as well.

We also performed an extensive numerical and simulation study to illustrate the applications of our results to familiar fading channel models. With some surprise, we have shown that homogenous threshold feedback policies are not always optimal to use for general fading channels, even when all MUs experience statistically the same channel conditions. In the particular case of Rayleigh fading channels, on the other hand, homogenous threshold feedback policies have been proven to be rate-wise optimal if multiple beams are used for the downlink communication. We have also studied the optimality and sub-optimality regions for the homogenous threshold feedback policies in the Rician and Nakagami case. The detailed insights regarding when and why homogenous threshold feedback policies are rate-wise optimal or suboptimal have been also provided.

Then, we focused the tradeoff between feedback and capacity in single-cell MIMO communication systems under $O(1)$ feedback constraints. Starting with an assumption of statistically identical MUs, we have obtained the tradeoff curve tracing the Pareto optimal boundary between feasible and infeasible feedback and capacity pairs. A point above the curve is unachievable, whereas all points below it are achievable, but sub-optimal in the sense that the same capacity scaling can be obtained by using strictly less feedback. We have provided the form of the homogeneous threshold feedback policies achieving the feedback-capacity pairs on the curve. We have also showed that if the $O(1)$ feedback constraint relaxed, we can achieve the same scaling achieved with full CSI with feedback load growing like $O\left((\log n)^{\epsilon}\right)$ for any $\epsilon \in (0,1)$.

We have also extended these results to a heterogeneous communication system, where different MUs experience non-identical path-loss gains. We have systematically altered the threshold levels at different MUs according to their large scale path-loss gains without violating the $O(1)$ feedback constraint. We showed that the heterogeneous network achieves the same performance in capacity scaling and feedback as the network consisting of statistically identical MUs. Since providing fairness among MUs is an important issue for such a heterogeneous communication environment, we have introduced two

new scheduling policies. We have showed that these scheduling policies coupled up with the systematic alteration of the threshold levels at different MUs allow the system to achieve optimum throughput scaling and fairness among MUs, simultaneously. Although most of these results are optimal in the limit for large networks, they also provide a close match with the asymptotically optimal results when used in finite size systems. In particular, we have showed that the threshold levels set by using the asymptotic formulas very rapidly achieve the required $O(1)$ feedback constraints in finite size systems as well.

Then, we made a subtle but interesting change to the optimization problem by setting an instantaneous/peak constraint on the number of MUs feeding back. Firstly, we defined a multiple access model which imposes this instantaneous feedback constraint on the system. The base station can reconstruct all the feedback packets successfully if and only if the random number of MUs feeding back is less than or equal to λ, which is the peak feedback constraint. We say that a collision occurs if the random number of MUs feeding back is greater than λ. In this case, all packets are destroyed together. We showed that this is a quasi-convex optimization problem by analyzing the rate expression, and solved it to obtain a formula for the optimal threshold value at the MU. The results do hold for most practical fading distributions, and we applied our results to the well known Rayleigh fading channel model to discuss the importance of the results. Finally, we analyzed the rate-wise optimality of using a homogenous threshold value. We provided a simple counter example which showed that even when considering a peak feedback constraint, a homogenous threshold level can be sub-optimal for s set of statistically identical MUs.

7.2 Future Work

Before concluding this thesis, we will briefly present some ideas which will motivate future studies on this topic.

7.2.1 Opportunistic Beamforming- Spatial Model With a Capped Gain

An introduction of a spatial model is not found in any of the analysis done on opportunistic beamforming. In almost all the analysis, the slow gain is neglected, which implies that all the users are equidistant from the base station. This issue can be addressed by introducing a spatial stochastic process governing user locations on top of the fast-fading model.

We can start with the system model presented in Subsection 2.2.1, *i.e.*, a single cell model in which M orthogonal beams are transmitted using N_t transmit antennas to n users. However, in addition, we can introduce a slow gain to take user locations into account. For user i, we denote the N_t-by-1 complex channel gain vector by $\boldsymbol{h}_i$. We call this the fast gain. $\sqrt{g_i}$, which is a random scalar, is the slow gain between the base station and user i. The signal received by the ith MU is given by

$$Y_i = \sqrt{\rho g_i} \sum_{k=1}^{M} \boldsymbol{h}_i^{\top} \boldsymbol{b}_k s_k + Z_i, \tag{7.1}$$

where ρ is the transmit power per beam, Z_i is the unit power (complex) Gaussian background noise, s_k and $\boldsymbol{b}_k$ are the transmitted symbol and the beamforming vector corresponding to the kth beam, respectively. Let $\gamma_{m,i}$ be the SINR value corresponding to the mth beam at the ith MU. Then, it is given by

$$\gamma_{m,i} = \frac{|\boldsymbol{h}_i^{\top} \boldsymbol{b}_m|^2}{(g_i \rho)^{-1} + \sum_{k=1, k \neq m}^{M} |\boldsymbol{h}_i^{\top} \boldsymbol{b}_k|^2}. \tag{7.2}$$

Now the SINRs on a beam among the n users are independent, but non-identically distributed random variables. We believe that this model can be successfully used in the future to analyze how the user locations would affect the results presented in this thesis.

7.2.2 Opportunistic Beamforming in a Multi-Cell Environment

Extending the results to a multi-cell environment will also be interesting. Let us consider a model comprising N identical base-stations/cells. There are n MUs in each cell, and each base station has N_t transmit antennas. Each MU is equipped with a single receive

antenna. $\boldsymbol{h}_{ij}$ is the N_t-by-1 multi antenna channel gain vector between user j and base station i. We assume that the elements of this vector are independent and identically distributed random variables. $\sqrt{g_{ki}}$ is the scalar slow gain from base station k to the users in cell i, and we assume that all the users in cell i have the same gain. Each base station transmits information along the directions of M orthonormal beamforming vectors. The transmitted signal from the ith base station can be written as

$$\boldsymbol{x}_i = \sqrt{\rho} \sum_{k=1}^{M} \boldsymbol{b}_{i,k} s_{i,k}, \tag{7.3}$$

where ρ is the transmit power per beam, and $s_{i,k}$ and $\boldsymbol{b}_{i,k}$ are the transmitted symbol and the beamforming vector corresponding to the kth beam from base station i, respectively. The signal received by the jth MU in cell i is given by

$$Y_{ij} = \sqrt{\rho g_{ii}} \boldsymbol{h}_{ij}^{\top} \boldsymbol{x}_i + \sum_{k \neq i}^{N} \sqrt{\rho g_{ki}} \boldsymbol{h}_{kj}^{\top} \boldsymbol{x}_k + Z_{i,j}, \tag{7.4}$$

where $Z_{i,j}$ is the unit power (complex) Gaussian background noise.

For simplicity, we can assume $0 < g_{ki} < 1$, and $g_{ki} = g_{ik}$, *i.e.*, a symmetric slow gain. If $N = 2$, $g_{i,i} = 1$ and $g_{i,k} = g_{k,i} = g$, this turns out to be the well known two cell Wyner model [100]. Using this model, we can search for optimal selective feedback policies for opportunistic beamforming in a multi-cell setting. Also, we can use this model to study the effects of base station cooperation in such an environment.

7.2.3 The Optimal Threshold Selection Problem

In Chapter 4, we have studied the optimal threshold selection problem for opportunistic beamforming. That is, given a finite feedback system with a feedback constraint λ, we studied the problem of determining the optimal threshold values which maximize the downlink throughput.

As mentioned in Chapter 4, this optimization problem is not easy to solve, even for a simple two-user system, due to the non-convex objective function, and a constraint set depending on the distribution of SINR values. Therefore, it is not possible to solve the

optimal threshold selection problem in its full generality for a general n-user system. In this thesis, we analyze a special case, which on itself is hard and interesting. The full case will be very challenging, and we believe it is an interesting problem to be addressed in the future.

Bibliography

[1] "Requirements related to technical performance for IMT-advanced radio interface(s)," ITU, Tech. Rep. M.2134, 2008.

[2] T. Al-Naffouri, M. Sharif, and B. Hassibi, "How much does transmit correlation affect the sum-rate scaling of MIMO Gaussian broadcast channels?" *IEEE Trans. Commun.*, vol. 57, pp. 562–572, Feb 2009.

[3] B. C. Arnold, "Majorization: Here, there and everywhere," *Statist. Sci.*, vol. 22, pp. 407–413, 2007.

[4] J. P. C. Avidor, D.; Ling, "Jointly opportunistic beamforming and scheduling (JOBS) for downlink packet access," in *Proc. International Conference on Communications*, vol. 5, Paris, France, Jun 2004, pp. 20–24.

[5] A. Bayesteh and A. Khandani, "On the user selection for MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 54, no. 3, pp. 1086 – 1107, Mar 2008.

[6] ——, "Asymptotic analysis of the amount of CSI feedback in MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1612 – 1629, Mar 2012.

[7] R. Bender and G. Sandstrom, "Wireless carriers refine 4G technology," *The Wall Street Journal*, Mar 2010.

[8] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and H. Poor, *MIMO Wireless Communications*. Cambridge University Press, 2007.

[9] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: Information-theoretic and communication aspects," *IEEE Trans. Inf. Theory*, vol. 44, pp. 2619–2692, Oct 1998.

[10] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Quantized vs. analog feedback for the mimo broadcast channel: A comparison between zero-forcing based achievable rates," in *Proc. IEEE International Symposium on Information Theory*, Nice, France, Jun 2007, pp. 2046–2050.

[11] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, Jul 2003.

[12] J. Caterer, "4G and next generation systems," *Information Technology Telecommunications and Electronics Association*, Feb 2009.

[13] G. Celik, G. Zussman, W. Khan, and E. Modiano, "MAC for networks with multipacket reception capability and spatially distributed nodes," *IEEE Trans. Mobile Comput.*, vol. 9, pp. 226–240, Feb 2010.

[14] D. Chizhik, J. Ling, P. W. Wolniansky, R. A. Valenzuela, N. Costa, and K. Huber, "Multiple-input-multiple-output measurements and modeling in Manhattan," *IEEE J. Sel. Areas Commun.*, vol. 21, pp. 321–331, Sep 2003.

[15] W. Choi, A. Forenza, J. Andrews, and R. Heath, "Opportunistic space-division multiple access with beam selection," *IEEE Trans. Commun.*, vol. 55, pp. 2371–2380, 2007.

[16] M. Costa, "Writing on dirty paper," *IEEE Trans. Inf. Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

[17] R. Couillet, J. Hoydis, and M. Debbah, "Random beamforming over quasi-static and fading channels: A deterministic equivalent approach," *To appear in IEEE Trans. Inf. Theory*, 2012.

[18] W. Dai, Y. Liu, B. Rider, and V. Lau, "On the information rate of MIMO systems with finite rate channel state feedback using beamforming and power on/off strategy," *IEEE Trans. Inf. Theory*, vol. 55, pp. 5032 – 5047, Nov 2009.

[19] H. A. David and H. N. Nagaraja, *Order Statistics*, 3rd ed. Wiley, 2003.

[20] J. Diaz, O. Simeone, and Y. Bar-Ness, "Sum-rate of MIMO broadcast channels with one bit feedback," in *Proc. IEEE International Symposium on Information Theory*, Seattle, WA, Jul 2006, pp. 1944–1948.

[21] ——, "Asymptotic analysis of reduced-feedback strategies for MIMO Gaussian broadcast channels," *IEEE Trans. Inf. Theory*, vol. 54, pp. 1308–1316, Mar 2008.

[22] S. N. Diggavi, N. Al-dhahir, A. Stamoulis, and A. R. Calderbank, "Great expectations: The value of spatial diversity in wireless networks," in *Proc. IEEE*, vol. 92, Feb 2004, pp. 219–270.

[23] Y. Eriguchi and T. Ohtsuki, "A random beamforming technique in multiuser multi-antenna OFDM systems for large system capacity and fairness among users," in *Proc. IEEE Vehicular Technology Conference*, Marina Bay, Singapore, May 2008, pp. 1–5.

[24] T. Eriksson and T. Ottosson, "Compression of feedback for adaptive transmission and scheduling," in *Proc.IEEE*, vol. 95, Feb 2007, pp. 2314–2321.

[25] ——, "Compression of feedback in adaptive ofdm-based systems using scheduling," *IEEE Commun. Letters*, vol. 11, pp. 859–861, 2007.

[26] E.Telatar, "Capacity of multi-antenna Gaussian channels," *European Transactions on Telecommunications*, vol. 10, pp. 585–598, Nov 1999.

[27] M. Fakhereddin, M. Sharif, and B. Hassibi, "Reduced feedback and random beamforming for OFDM MIMO broadcast channels," *IEEE Trans. Commun.*, vol. 57, pp. 3827 – 3835, Dec 2009.

[28] F. Floren, O. Edfors, and B.-A. Molin, "The effect of feedback quantization on the throughput of a multiuser diversity scheme," in *Proc. IEEE Global Telecommunications Conference*, vol. 1, San Francisco, USA, Dec 2003, pp. 497–501.

[29] D. Gesbert and M. Alouini, "How much feedback is multi-user diversity really worth?" in *Proc. International Conference on Communications*, Paris, France, Jun 2004, pp. 234–238.

[30] D. Gesbert, M. Kountouris, R. Heath, C. Chan-Byoung, and T. Salzer, "Shifting the MIMO paradigm," *IEEE Signal Process. Mag.*, vol. 24, pp. 36–46, 2007.

[31] S. Ghez, S.Verdu, and S. Schwartz, "Stability properties of slotted aloha with multipacket reception capability," *IEEE Trans. Auto. Control*, vol. 33, pp. 640–649, Jul 1988.

[32] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products Seventh Edition*. Academic Press, 2007.

[33] A. Hadden, "Mobile broadband - where the next generation leads us [industry perspectives]," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 6–9, Dec 2009.

[34] V. Hassel, D.Gesbert, M.-S. Alouini, and G. Oien, "A threshold-based channel state feedback algorithm for modern cellular systems," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 2422–2426, Jul 2007.

[35] R. Heath-Jr. and A. Paulraj, "A simple scheme for transmit diversity using partial channel feedback," in *Proc. Asilomar Conference on Signals, Systems and Computers*, vol. 2, CA, USA, Nov 1998, pp. 1073–1078.

[36] Y. Huang and B. Rao, "Closed form sum rate of random beamforming," *IEEE Commun. Letters*, vol. 16, pp. 630 – 633, May 2011.

[37] H. Inaltekin, T. Samarasinghe, and J. Evans, "Rate optimal limited feedback policies for the MIMO downlink," in *Proc. International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, Princeton, USA, May 2011.

[38] V. Jimenez and A. Armada, "An adaptive mimo - ofdm system: Design and performance evaluation," in *Proc.IEEE Int. Symposium on Wireless Communication Systems*, Valencia, Spain, Sep 2006, pp. 809–813.

[39] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5045–5060, Nov 2006.

[40] C. Kim, S. Jung, and J. Lee, "Post-scheduling SINR mismatch analysis for multiuser orthogonal random beamforming systems," *IEEE Trans. Commun.*, vol. 59, pp. 1509–1513, Jun 2011.

[41] I.-M. Kim, S.-C. Hong, S.S.Ghassemzadeh, and V.Tarokh, "Opportunistic beamforming based on multiple weighting vectors," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 2683–2687, Nov 2005.

[42] R. Knopp and P. Humblet, "Information capacity and power control in single cell multiuser communications," *Proc. International Conference on Communications, Seattle, USA*, 1995.

[43] V. Lau, L. Youjian, and T.-A. Chen, "On the design of mimo block-fading channels with feedback-link capacity constraint," *IEEE Trans. Commun.*, vol. 52, pp. 62–70, 2004.

[44] J. Lei, Y. Li, X. Chen, S. Zhou, J. Wang, and Y. Yao, "Cooperative opportunistic scheduling in multiple antenna cellular networks," in *Proc. IEEE Vehicular Technology Conference*, Barcelona, Spain, Apr 2009, pp. 1–5.

[45] Z. Li, J. Yang, and J. Yao, "Double-threshold opportunistic transmission strategy for MU-MIMO downlink," in *Proc. International Symposium on Communications and Information Thechnology*, Incheon, Korea, Sep 2009, pp. 319–323.

[46] Y. Liu, L. Zhang, and G. Lu, "Joint optimization for PAPR reduction and opportunistic beamforming (OBF) in OFDM systems," in *Proc.International Conference on Wireless Communications, Networking and Mobile Computing*, Beijing, China, Sep 2009, pp. 1–4.

[47] D. Love and R. Heath, "Limited feedback unitary precoding for spatial multiplexing systems," *IEEE Trans. Inf. Theory*, vol. 51, no. 08, pp. 2967 – 2976, Aug 2005.

[48] D. Love, R. Heath, V. Lau, D. Gesbert, and B. R. M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 26, pp. 1341–1365, Oct 2008.

[49] D. G. Luenberger, *Optimization by Vector Space Methods*. New York, NY: Wiley, 1968.

[50] N. Mahravari, "Random-access communication with multiple reception," *IEEE Trans. Inf. Theory*, vol. 36, pp. 614–622, May 1990.

[51] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*. New York, NY: Academic Press, 1979.

[52] P. Mathys, "A class of codes for a t active users out of n multiple-access communication system," *IEEE Trans. Inf. Theory*, vol. 36, pp. 1206–1219, Nov 1990.

[53] L. Min and F. Guangzeng, "User selection for multiuser MIMO BC with zero-forcing beamforming," in *Proc.International Conference on Wireless Communications Networking and Mobile Computing*, Chengdu, China, Sep 2010, pp. 1–5.

[54] M.Kountouris and D.Gesbert, "Memory-based opportunistic multi-user beamforming," in *Proc. IEEE International Symposium on Information Theory, Adelaide, Australia*, Sep 2005, pp. 1426–1430.

[55] ——, "Robust multi-user opportunistic beamforming for sparse networks," *IEEE Workshop on Signal Processing Advances in Wireless Communications*, pp. 975–979, Jun 2005.

[56] M.Kountouris, D.Gesbert, and T.Salzer, "Enhanced multiuser random beamforming: Dealing with the not so large number of users case," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1536–1545, Oct 2008.

[57] S. Moon, S. Lee, and I. Lee, "Sum-rate capacity of random beamforming for multi-antenna broadcast channels with other cell interference," *IEEE Trans. Wireless Commun.*, vol. 10, pp. 2440–2444, Aug 2011.

[58] S.-H. Moon, S.-R. Lee, and I. Lee, "Sum-rate capacity of random beamforming for multi-antenna broadcast channels with other cell interference," *IEEE Trans. Wireless Commun.*, vol. 10, pp. 2440 – 2444, Aug 2011.

[59] M.Vemula, D.Avidor, J.Ling, and C.Papadias, "Inter-cell coordination, opportunistic beamforming and scheduling," in *Proc. International Conference on Communications*, vol. 12, Istanbul, Turkey, Jun 2006, pp. 5319–5324.

[60] B. L. Ng, J. S. Evans, S. V. Hanly, and D. Aktas, "Transmit beamforming with cooperative base stations," in *Proc. IEEE International Symposium on Information Theory*, Adelaide, Australia, Sep 2005, pp. 1431–1435.

[61] H. Nguyen and T. Lestable, "Compression of bit loading power vectors for adaptive multi-carrier systems," in *Proc.IEEE Int. Midwest Symposium on Circuits and Systems*, vol. 3, Hiroshima, Japan, Jul 2004, pp. 243–246.

[62] H. Nguyen, R. Zhang, and H. Hui, "Multi-cell random beamforming: Achievable rate and degrees of freedom region," *Submitted to IEEE Trans. Signal Process.*, 2012.

[63] V. K. Nguyen and J. S. Evans, "Multiuser transmit beamforming via regularized channel inversion: A large system analysis," in *Proc. IEEE Global Telecommunications Conference*, New Orleans, USA, Nov 2008, pp. 1–4.

[64] M. Nicolaou, A. Doufexi, and S. Armour, "Reducing feedback requirements of the multiple weight opportunistic beamforming scheme via selective multiuser diversity," in *Proc. IEEE Vehicular Technology Conference*, Marina Bay, Singapore, May 2008, pp. 1–5.

[65] O. Ozdemir and M. Torlak, "Performance of opportunistic beamforming with quantized feedback," in *Proc. IEEE Global Telecommunications Conference, San Francisco, USA*, Dec 2006, pp. 1–5.

[66] Z. Pan and L. Chen, "A novel simplified opportunistic beamforming method for wide-band systems," in *Proc.Wireless Communications and Networking Conference*, Las Vegas, USA, Apr 2006, pp. 1747–1752.

[67] L. Pang and M. Jiang, "Sum rate capacity for orthogonal random beamforming in multi-cell MIMO scenarios," in *Proc.World Congress on Intelligent Control and Automation*, Shandong, China, Jul 2010, pp. 1440 – 1443.

[68] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. McGraw Hill, 1991.

[69] K.-H. Park, Y.-C. Ko, and M. Alouini, "Accurate approximations and asymptotic results for the sum-rate of random beamforming for multi-antenna Gaussian broadcast channels," in *Proc. IEEE Vehicular Technology Conference*, Barcelona, Spain, Apr 2009, pp. 1–6.

[70] M. Pugh and B. D. Rao, "Reduced feedback schemes using random beamforming in MIMO broadcast channels," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1821–1832, Mar 2010.

[71] X. Qin and R. Berry, "Distributed approaches for exploiting multiuser diversity in wireless networks," *IEEE Trans. Inf. Theory*, vol. 52, pp. 392–413, Feb 2006.

[72] L. Roberts, "ALOHA packet system with and without slots and capture," *ACM SIGCOMM Computer Communication Review*, vol. 5, pp. 28–42, Apr 1975.

[73] T. Samarasinghe, H. Inaltekin, and J. Evans, "The feedback-capacity tradeoff for opportunistic beamforming," in *Proc. International Conference on Communications*, Kyoto, Japan, Jun 2011.

[74] ——, "Optimal selective feedback policies for opportunistic beamforming," Submitted to IEEE Trans. Inf. Theory, Available: http://arxiv.org/abs/1107.4822, 2011.

[75] ——, "Vector broadcast channels: Optimal threshold selection problem," in *Proc. IEEE International Symposium on Information Theory*, St. Petersburg, Russia, Aug 2011.

[76] ——, "Vector broadcast channels: Optimality of threshold feedback policies," in *Proc. IEEE International Symposium on Information Theory*, St. Petersburg, Russia, Aug 2011.

[77] ——, "The feedback-capacity tradeoff for opportunistic beamforming under optimal user selection," *Performance Evaluation*, 2012.

[78] S. Sanayei and A. Nosratinia, "Opportunistic beamforming with limited feedback," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 2765–2771, Aug 2007.

[79] M. Shariff and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 506–522, Feb 2005.

[80] ——, "A comparison of time-sharing, DPC, and beamforming for MIMO broadcast channels with many users," *IEEE Trans. Commun.*, vol. 55, no. 1, pp. 11–15, Jan 2007.

[81] N. Shrivastava, S. Singh, and A. Trivedi, "Joint scheduling and random beamforming with reduced feedback in multiuser MIMO-OFDM," in *Proc.Wireless Advanced*, London, UK, Jun 2011, pp. 66–69.

[82] B. Sklar, "Rayleigh fading channels in mobile digital communication systems part i: Characterization," *IEEE Commun. Mag.*, vol. 35, pp. 136–146, Sep 1997.

[83] ——, "Rayleigh fading channels in mobile digital communication systems part ii: Mitigation," *IEEE Commun. Mag.*, vol. 35, pp. 148–155, Sep 1997.

[84] H. Son and S. Lee, "Iterative best beam selection for random unitary beamforming," *IEEE Trans. Commun.*, vol. 59, pp. 968–974, Jun 2011.

[85] Q. Spencer, C. Peel, A. Swindlehurst, and M. Haardt, "An introduction to the multi-user MIMO downlink," *IEEE Commun. Mag.*, vol. 42, pp. 60–67, Oct 2004.

[86] Q. Spencer and A. Swindlehurst, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Process.*, vol. 52, pp. 461 – 471, Feb 2004.

[87] Y. Suchiya, T. Ohtsuki, and T. Kaneko, "Random beamforming using low feedback and low latency power allocation," in *Proc. IEEE Vehicular Technology Conference*, Dublin, Ireland, Oct 2007, pp. 402–406.

[88] X. Tang, S. A. Ramprashad, and H. Papadopoulos, "Multi-cell user-scheduling and random beamforming strategies for downlink wireless communications," in *Proc. IEEE Vehicular Technology Conference*, Barcelona, Spain, Apr 2009, pp. 1–5.

[89] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Trans. Inf. Theory*, vol. 44, pp. 744–765, Mar 1998.

[90] S. Thoen, L. V. der Perre, B. Gyselinckx, and M. Engels, "Performance analysis of combined transmit-sc/receive-mrc," *IEEE Trans. Commun.*, vol. 49, pp. 5–8, 2001.

[91] D. N. C. Tse and P. Viswanath, *Fundamentals of Wireless Communications*. Cambridge University Press, 2005.

[92] J. Vicario, R. Bosisio, and U. Spagnolini, "Adaptive beam selection techniques for opportunistic beamforming," in *Proc.International Symposium on Personal, Indoor and Mobile Radio Communications*, Helsinki, Finland, Sep 2006, pp. 1–5.

[93] J. Vicario, R. Bosisio, U. Spagnolini, and C. Anton-Haro, "A throughput analysis for opportunistic beamforming with quantized feedback," in *Proc.International Symposium on Personal, Indoor and Mobile Radio Communications*, Helsinki, Finland, Sep 2006, pp. 1–5.

[94] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2658–2668, Oct 2003.

[95] P. Viswanath, V. Anantharam, and D. N. C. Tse, "Optimal sequences, power control, and user capacity of synchronous CDMA systems with linear MMSE multiuser receivers," *IEEE Trans. Inf. Theory*, vol. 45, pp. 1968–1983, Sep 1999.

[96] P. Viswanath and D. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1912–1921, Aug 2003.

[97] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1277–1294, Jun 2002.

[98] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian MIMO broadcast channel," in *Proc. IEEE International Symposium on Information Theory*, Chicago, USA, Jun 2004.

[99] J.-H. Wu, Y.-T. Hsieh, and P. Ting, "Multicell collaborative random beamforming with power sharing and user selection," in *Proc. IEEE Global Telecommunications Conference*, Miami, USA, Dec 2010, pp. 1–5.

[100] A. D. Wyner, "Shannon-theoretic approach to a Gaussian cellular multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 40, pp. 1713–1727, Nov. 1994.

[101] M. Xia, Y. Zhou, J.Ha, and H. K. Chung, "Opportunistic beamforming communication with throughput analysis using asymptotic approach," *IEEE Trans. Veh. Technol.*, vol. 58, no. 5, pp. 2608–2614, Jun 2009.

[102] Y. Xu, J. Zhao, and Y. Cai, "User selection based on feedback threshold for MIMO broadcast channels," in *Proc. International Conference on Signal Processing and Communication Systems, Gold Coast, Australia*, Dec 2008.

[103] Y. Xue and T. Kaiser, "Exploiting multiuser diversity with imperfect one-bit channel state feedback," *IEEE Trans. Veh. Technol.*, vol. 56, no. 1, pp. 183–193, Jan 2007.

[104] H. Yang, P. Lu, H. Sung, and Y. Ko, "Exact sum-rate analysis of mimo broadcast channels with random unitary beamforming," *IEEE Trans. Commun.*, vol. 59, pp. 2982–2986, Nov 2011.

[105] J. Yang, Y. Kim, and D. Kim, "A closed form solution for the sum rate of multiple random beamforming," *IEICE Trans. Commun.*, vol. E92-B, no. 4, pp. 1373–1375, Apr 2009.

[106] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 528–541, Mar 2006.

[107] Y. J. Zhang, "Multi-round contention in wireless LANs with multipacket reception," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 1503–1513, Apr 2010.

[108] Y. J. Zhang, P. X. Zheng, and S. C. Liew, "How does multiple-packet reception capability scale the performance of wireless local area networks?" *IEEE Trans. Mobile Comput.*, vol. 8, pp. 923–935, Jul 2009.

[109] P. X. Zheng, Y. J. Zhang, and S. C. Liew, "Multipacket reception in wireless local area networks," in *Proc. International Conference on Communications*, vol. 8, Istanbul, Turkey, Jun 2006, pp. 3670–3675.