ASYMPTOTIC PERFORMANCE ANALYSIS AND TRANSMISSION DESIGN FOR

LARGE-SCALE RELAY NETWORKS

by

Qian Wang

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Communications

Department of Electrical and Computer Engineering
University of Alberta

Abstract

The large-scale relay network is a promising component of future wireless systems, as it can improve the coverage and throughput of wireless communications. The large-scale scenario brings many new challenges. In this thesis, two of the fundamental problems are investigated, i.e., performance analysis and transmission design for large-scale relay networks.

For performance analysis, this thesis studies both distributed relaying schemes, e.g., relay selection and distributed relay beamforming, and centralized relaying scheme, e.g., maximal-ratio combining/transmission (MRC/MRT). For best relay selection (BRS), closed-form expressions of the average received SNR and ergodic capacity are derived, which provide insights on the array gain and ergodic capacity behaviour of BRS. For distributed relay beamforming, the power allocation that maximizes the sum-rate is proposed. Then the asymptotic behaviour of the SNR is derived rigorously for the high transmit power regime. For MRC/MRT, a comprehensive performance scaling law and performance analysis is provided in a multi-user massive MIMO relay network with channel state information (CSI) error. The results show quantitatively the trade-off between the network parameters and their effects on the performance. In addition, a sufficient condition on the parameter scalings for the signal-to-interference-plus-noise-ratio (SINR) to be asymptotically deterministic is derived, which covers existing studies as special cases. Further, the scenario where the SINR increases linearly with the number of relay antennas is studied. The sufficient and necessary condition on the parameter scaling for this

scenario is proved. The outage probability and average bit error rate (ABER) of the relay network in this case are analysed. Besides performance analysis, rank detection design is also investigated. Due to the spatial correlation and antenna coupling, the large-scale channel matrix usually has reduced rank. Accurate rank detection is crucial in the estimation of such channel matrices. Several rank detection methods are proposed, which provides higher rank detection rate than existing ones.

$\sim$

Acknowledgement

I hope to give my greatest gratitude and respect to my supervisor, Dr. Yindi Jing for her efforts through the years in helping me to be a better researcher, and more importantly, a better man. During the long-term research period, she guided me with her rich knowledge and experience, taught me the way of logical thinking and writing, which not only boosted my research, but also will help me achieve more in the rest of my life. She is not only a supervisor to me, but also a good friend, who would like to listen to my trouble and share her experience in life. I greatly appreciate it for all her help.

I also want to thank the committee members of my PhD exam and candidacy exam, Dr. Hai Jiang, Dr. Chintha Tellambura, Dr. Vicky Zhao, Dr. Qing Zhao, Dr. Majid Khabbazian, Dr. Mahdi Tavakoli, and Dr. Xiaodai Dong for their time reviewing my research and their valuable suggestions. My thanks also go to the faculty and staff in the department, who provide great support for my study.

Specially thanks to my fellow lab-mates for their help with my research and all the happy time with them.

I also should not forget to thank the China Scholarship Council for the financial support.

At last, I give my deepest gratitude to my family who have supported me all the way. Specially thanks to my wife, Wei Wang, and my son, Will Wang. They made my Ph.D. life filled with happiness and love.

$$\sim$$

Table of Contents

List of Tables

List of Figures

List of Symbols

Elementary & Special Functions

Notation	Definition
$\Gamma(\cdot)$	Gamma function
$\gamma(\cdot, \cdot)$	lower incomplete gamma function
$\Gamma(\cdot, \cdot)$	upper incomplete gamma function
$K_\nu(\cdot)$	the ν-th order modified Bessel function of the second kind
$\exp(\cdot)$	exponential function
$W_{-1}(\cdot)$	Lambert W function
$\mathrm{erfc}(\cdot)$	complementary error function
$Q(\cdot)$	Q function, $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\left(-\frac{u^2}{2}\right) du$
$\lfloor \cdot \rfloor$	floor function
$\log_2, \log_{10}, \ln$	logarithm function with base 2, 10 or Euler number e

Probability & Statistics

Notation	Definition
$\mathbb{E}\{X\}$	expected value of X
$\mathrm{Var}\{X\}$	variance of X
$\mathrm{Cov}\{X, Y\}$	covariance between X and Y
$\mathrm{SCV}\{X\}$	squared coefficient of variation of X
$X \sim \mathcal{CN}(m, \sigma^2)$	circular-symmetric complex Gaussian random variable X with mean m and variance σ^2

Miscellaneous

Notation	Definition
$\mathbb{C}$	set of complex numbers
$|\cdot|$	modulus of a complex number

Notation	Definition
$\max_{n=1,2,\ldots,N}\{a_n\}$	maximum of scalars a_n, $n \in \{1,2,\ldots,N\}$
$\min_{n=1,2,\ldots,N}\{a_n\}$	minimum of scalars a_n, $n \in \{1,2,\ldots,N\}$
$\lim_{x \to a} f(x)$	the limit of function $f(x)$ as x tends to a
$f(x) = \mathcal{O}(g(x))$	$f(x)$ and $g(x)$ are two functions of x defined on the subset of real numbers, this expression means that when $x \to \infty$, there is a positive constant C such that for all sufficiently large values of x, the absolute value of $f(x)$ is at most C multiplied by the absolute value of $g(x)$
$f'(x), \mathrm{d}f(x)/\mathrm{d}x$	the derivative of function $f(x)$ with respect to x
$\mathbf{A}^*$	Hermitian of matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	Inverse of the square matrix $\mathbf{A}$
$\mathrm{tr}(\mathbf{A})$	trace of matrix $\mathbf{A}$
$\mathrm{rank}(\mathbf{A})$	rank of matrix $\mathbf{A}$
$\det(\mathbf{A})$	determinant of matrix $\mathbf{A}$
$\mathbf{I}_n$	$n \times n$ identity matrix
$\mathrm{diag}\{a_1,\ldots,a_n\}$	diagonal matrix whose diagonal entries starting in the upper left corner are $a_1, \cdots, a_n$
$\emptyset$	empty set
$\xrightarrow{d}; \xrightarrow{P}; \xrightarrow{m.s.}$	convergence in distribution; convergence in probability; almost surely convergence
$a \ll b; a \gg b$	a is much less than b; a is much larger than b

List of Abbreviations

Abbreviation	Definition
1G/2G/../5G	1st/2nd/../5th generation of wireless systems
3D	Three-dimensional
AF	Amplify-and-forward
ABER	Average bit error rate
BRS	Best relay selection
BER	Bit error rate
CDF	Cumulative distribution function
CLT	Central limit theorem
CoMP	Coordinated multipoint
CSCG	Circular-symmetric complex Gaussian
CSI	Channel state information
D2D	Device-to-device
DF	Decode-and-forward
EDGE	Enhanced data rates for GSM evolution
EVT	Extreme value theory
FDMA	Frequency division multiple access
GPRS	General packet radio service
GSM	Global system for mobile communications
IEEE	Institute of electrical and electronics engineers
IMT-Advanced	International mobile telecommunications advanced
ITU-R	International telecommunication union-radio
i.i.d.	Independent and identically distributed
LTE	Long-term evolution
MIMO	Multiple-input multiple-output
ML	Maximal likelihood
MMSE	Minimum mean squared error
MRC	Maximal ratio combining
MRT	Maximal ratio transmission
MSE	Mean squared error
PA	Power allocation
PDF	Probability density function
QoS	Quality-of-service
RF	Radio-frequency

Abbreviation	Definition
RRU	Remote radio unit
SCV	Squared coefficient of variation
SINR	Signal-to-interference-plus-noise-ratio
SNR	Signal-to-noise-ratio
SVD	Singular value decomposition
SURE	Stein unbiased risk estimate
TDD	Time-division-duplexing
TDMA	Time division multiple access
WiMAX	Worldwide interoperability for microwave access
ZF	Zero-forcing

Chapter 1

Introduction

The ultimate goal of wireless networks is to enable communication of any type of information with anyone, at anytime, from anywhere. We are getting closer and closer to this goal with the ever-evolving wireless technologies. It started from 1980's, when the first generation of wireless systems (1G), the analog cellular system, was deployed. At the time, users could make voice calls within one country. Since then, the wireless technology has found itself significantly improved in every decade. In 1990's, the second generation of wireless systems (2G) (e.g., Global System for Mobile Communications (GSM), Interim Standard-95, etc.) was launched. With digital signal processing applied, not only the voice transmission rate had been increased more than 10-fold of 1G, but also the data services were introduced for the first time, including text images and multi-media. The maximum data rate of 2G is 50 kbit/s with General Packet Radio Service (GPRS), and 500 kbit/s with Enhanced Data Rates for GSM Evolution (EDGE). While 2G is mainly built for voice services and slow data transmission, the third generation of wireless systems (3G), developed in late 1990's, was intended for high-volume data services. It provides a minimum data rate of 2 Mbit/s for stationary or walking users, and 384 kbit/s in a moving vehicle. The data rate improvement enables various data-driven services, e.g., mobile TV, video conferencing, and on-line gaming. Driven by even data-desiring services, e.g., high-definition mobile TV and three-dimensional (3D) television, and the booming mobile devices, e.g., smart phones, laptops, and tablets, the evolution of the wireless networks came to the fourth generation of wireless systems (4G). In March 2008, the International Telecommunication Union-Radio

(ITU-R) communication sector specified a set of requirements for 4G standards, named the International Mobile Telecommunications Advanced (IMT-Advanced) specification, setting peak rate of 4G at 100 Mbit/s for high mobility communications and 1 Gbit/s for low mobility communications.

The above is the evolution of wireless communication technologies in the past decades. The driving force behind the innovations is the booming user demand. On one hand, the number of mobile devices increases dramatically. By 2015, global mobile devices and connections have grown up to 7.9 billion, with a 563 million increase in 2015 [1]. On the other hand, the mobile services have expanded from voice call services to the diverse mobile video and social networking services, which leads to more data usage per mobile device. These result in huge data traffic growth. In the past ten years, the mobile data traffic has grown 4,000-fold (Fig. 1.1). In 2015 alone, the global mobile data traffic grew 74 percent [1]. This trend will continue even more wildly in the next few years. By 2020, the mobile data traffic will be almost 10 times of 2015 (Fig. 1.2). This urges further evolution of the wireless technology, and calls for the the fifth generation of wireless systems (5G).
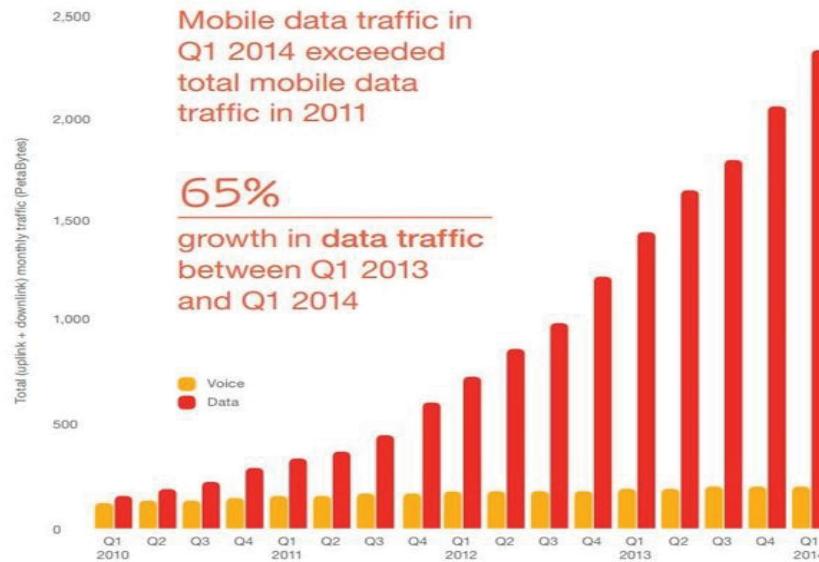
Figure 1.1: Global mobile data traffic growth [2].

2

Figure 2. Cisco Forecasts 30.6 Exabytes per Month of Mobile Data Traffic by 2020

53% CAGR 2015–2020

Exabytes per Month

- 3.7 EB (2015)
- 6.2 EB (2016)
- 9.9 EB (2017)
- 14.9 EB (2018)
- 21.7 EB (2019)
- 30.6 EB (2020)
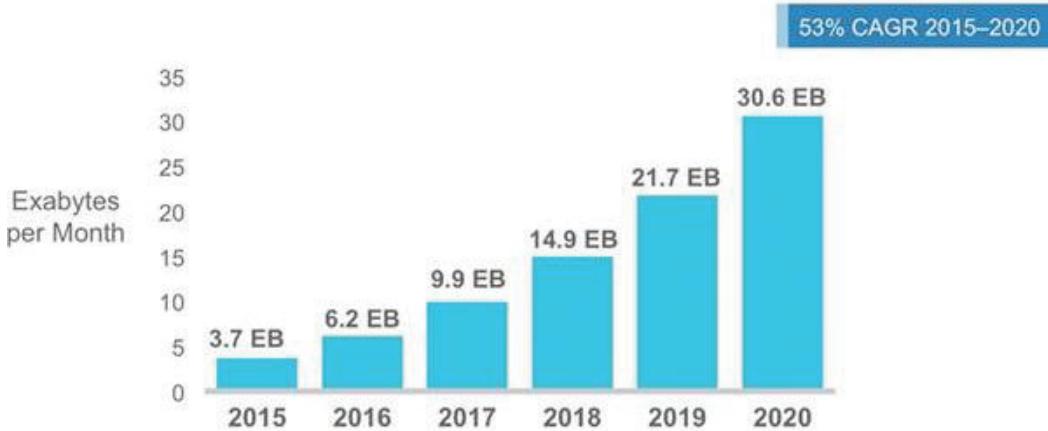
Source: Cisco VNI Mobile, 2016

Figure 1.2: Forecast of global mobile data traffic growth [1].

To accommodate the huge data traffic, 5G must support higher capacity with better reliability, higher spectrum and energy efficiency. One technology that is promising to meet these requirements is massive MIMO, where a large number of antennas are deployed at the base station. 5G should also provide ubiquitous coverage, i.e., the service should be available to anyone from anywhere, even for users in rural areas or at the cell edge of a cellular network. One technology that extends the coverage of a wireless network is the relay technology, which assists the communication between nodes that do not have reliable direct connection. The combination of massive MIMO and relay is called the large-scale relay network. It is expected to inherit the advantages of both and play an important role in 5G [1].

1.1 Large-Scale Relay Network

The concept of large-scale relay network originates from the conventional relay technology, and incorporates the ascendant massive MIMO technology. In what follows, the two important components will be introduced.

---

[1] Other famous technologies for 5G include cognitive radio, millimetre wave, and etc. Engaged readers can find their introductions in [7]

1.1.1 Relay Communications

The basic idea of relay is to forward messages from sources to destinations. Originated since 1980's, relay technology has witnessed a rapid growth in the past decades, which has been integrated into various 3G and 4G standards [3–5] and being a strong candidate for 5G [6]. In the conceptual 5G hybrid topology (Figure. 1.3) [7], relay finds its diverse applications, e.g., mobile relay, multi-hop relay, and user-equipment-based relay[2]. In the following, the benefits of relays in future wireless networks are generally summarized.
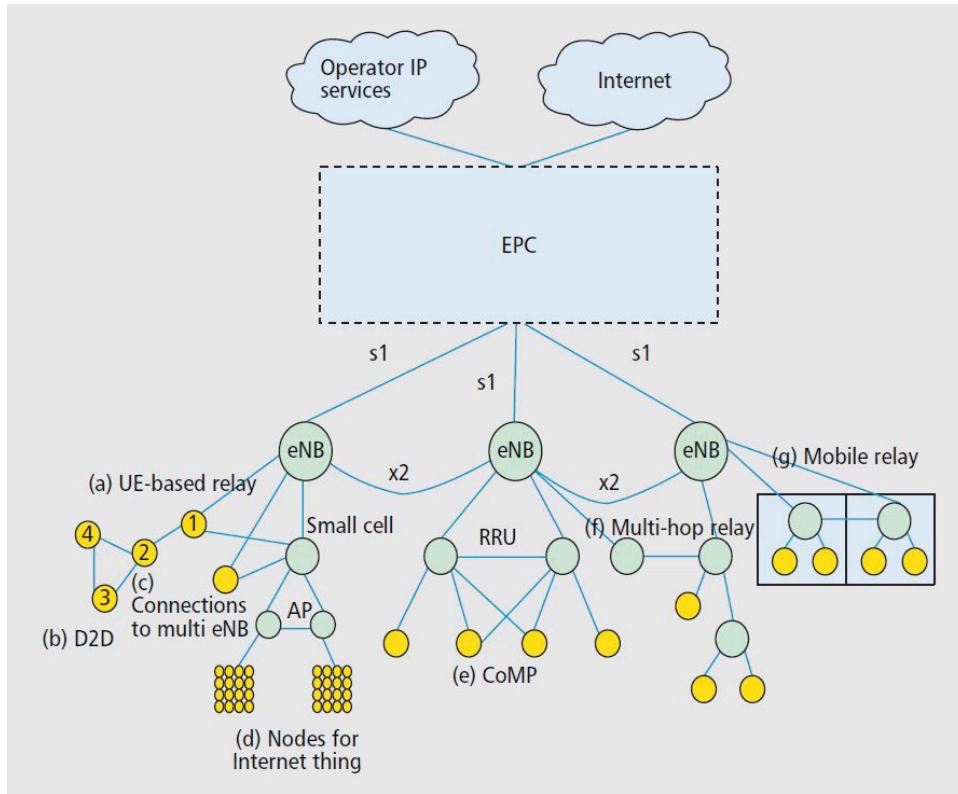
Figure 1.3: Current vision of the hybrid networking topology in 5G [7]. In the figure, EPC stands for Evolved Packed Core, which is the core network. eNB stands for Evolved NodeB, which is the base station.

- **Improving cell-edge performance.** In cellular systems, user equipments at the cell edge experience severe path-loss because of the long distance from

---

[2]Note that, besides relaying, there are various advanced technologies applied in this network, e.g., device-to-device (D2D), small cell, coordinated multipoint (CoMP), and remote radio unit (RRU). Engaged readers can find their introductions in [7].

the base station. To guarantee the quality-of-service (QoS) of those users, traditionally, very high transmit power is used to overcome the path-loss. In the hybrid-topology cellular network, one relay or multi-hop relay stations can be deployed between the base station and the cell-edge users [7]. The use of relay will reduce the path-loss, extend the coverage of the cellular network, and also reduce the transmit power of both the base station and users [8].

- **Assisting mobile communications.** User equipments in a mobile vehicle, e.g., a mobile train, experience severe penetration-loss when the signal penetrates walls. Implementing a relay station on the mobile vehicle can eliminate the penetration-loss and improve the communication quality of mobile users [7]. Moreover, the transmit power of user devices can be reduced significantly as the relay is much closer than the base station.

- **Assisting D2D communications.** The D2D communication is an important component of the hybrid topology of future wireless networks [7]. In D2D systems, devices communicate with each other directly without traversing the base station or the core network. The application of D2D releases the pressure on the core network and improves the energy efficiency and spectral efficiency. When there is no reliable direct link between devices, a relay station can be deployed between source devices and destination devices [9]. Or alternatively, the devices between sources and destinations can act as relays and cooperate to assist the communications [8]. In either way, the capacity of D2D systems will be improved.

These are the advantages of relay technology in future wireless networks. To further boost its performance for new 5G applications, the future relay technology will incorporate another promising technology, massive MIMO, which will be introduced next.

1.1.2 Massive MIMO

Multiple-input multiple-output (MIMO) systems, or multiple antenna systems, refer to systems with multiple antennas implemented at the transceiver nodes. The

MIMO technology has matured and has been incorporated into many wireless standards including the Institute of Electrical and Electronics Engineers (IEEE) $802.11n$, IEEE $802.11ac$, Worldwide Interoperability for Microwave Access(WiMAX), and Long-Term Evolution (LTE) [29, 30]. For example, the LTE standard allows for up to 8 antenna ports. With more antennas at the transmitter/receiver, the propagation channel will have more degrees of freedom, which can be exploited to provide higher data rate or link reliability to fulfil the huge demand of data rate in future wireless networks. Besides, it is more practical and efficient compared with other solutions, e.g., increasing frequency spectrum or transmission power, considering that the available frequency spectrum is limited and the future communications need to be more power efficient. Therefore, a novel scheme, called large-scale MIMO or massive MIMO, where hundreds of antennas are implemented at the transmitter nodes [10], has attracted much attention recently. With the massive MIMO technology, the following advantages can be gained over conventional MIMO systems.

- **Higher energy efficiency.** While conventional MIMO is energy efficient, with a large number of antennas, energy can be focused with extreme sharpness into the small regions of desired terminals, making massive MIMO more efficient in energy usage.

- **Reduced effect of small-scale fading.** Small-scale fading is a fundamental feature of wireless propagation that conventional MIMO systems exploit. With very large scale, the randomness of channel vectors can be reduced by using simple processing techniques. Thus the effect of small-scale fading can be asymptotically eliminated in massive MIMO systems.

- **Reduced interference.** The inter-user interference is a major problem in MIMO systems with multiple users. With massive MIMO, due to the large scale, the random channel vectors between different users and the base stations become mutually orthogonal asymptotically. Thus, the interference can be asymptotically cancelled with simple processing.

With these advantages, the massive MIMO concept is widely recognized to be a basic component of 5G networks [7].

6

Massive MIMO technology has the following application scenarios: 1) multi-user networks where a massive MIMO base station serves many user devices which are equipped with limited number of antennas; 2) relay networks with a massive MIMO relay station or a large number of distributed single-antenna relays help the communications from a group of user devices to another group of user devices. The latter case is called the large-scale relay network. Compared with conventional relay networks, the large-scale relay network can not only improve the network coverage, but also provide much higher throughput and serve more users, to meet the requirement of future wireless networks. This thesis focuses on the research of large-scale relay networks.

1.2 Thesis Motivation and Contributions

1.2.1 Motivation

As mentioned before, the large-scale relay network is a strong candidate for future wireless networks for its potential to increase the throughput, coverage, energy efficiency and spectrum efficiency. The brief procedure of communication in a large-scale relay network is introduced as follows. Firstly, the signals are transmitted from the sources to the relays. Then the relays process the signals and retransmit to the destinations. The processing at the relays is called the relaying schemes. Popular relaying schemes include relay selection, distributed relay beamforming, maximal-ratio combining/transmission (MRC/MRT), and so on. Relaying schemes are the key to the high performance of the relay network. On the other hand, as most relaying schemes require channel state information (CSI) at the relay, channel estimation is conducted before the data transmission.

From the procedure, we can see that the performance analysis of relaying schemes and transmission design play important roles in understanding and designing the network. The two problems will be described in detail as below.

- **Performance analysis.** Performance analysis shows the properties of different relaying schemes and effects of the network parameters on the performance. These results will help to design the large-scale relay networks in

practice.

In this thesis, we are looking for neat, systematic, general, closed-form or semi-closed-form expressions for the performance of large-scale relay networks. From the expressions, the network performance with respect to different parameters can be easily observed. This problem is challenging and different from conventional relay performance analysis for the following reasons.

- The analysis of multiple-antenna relay networks has always been a challenging problem. Existing performance analyses for traditional relay networks [11–18] are based on the distribution of the signal-to-noise-ratio (SNR). However, due to the complexity of the two-hop transmission, i.e., the source-to-relay and relay-to-destination transmissions, explicit expressions of performance metrics, e.g., average SNR and ergodic capacity, are rare and little insights are available.

- In large-scale relay networks, it is of more importance to derive the asymptotic performance behaviour of the network when the relay antenna number is very large. This is different to conventional relay networks, where the performance analysis mainly focuses on the limited relay antenna number case.

- **Transmission design.** For transmission design, we specialise on the rank detection problem. The channel matrix of large-scale relay networks usually has reduced rank [10]. For reduced-rank channel, accurate rank detection is important for the following two reasons. On one hand, it can improve the channel estimation quality, which is crucial for the signal processing at the relay. On the other hand, when the relay is serving multiple users, the rank of channel matrix between users and the relay determines how many users can be served within the same time-frequency bandwidth. Thus, accurate channel rank detection is an important part of channel estimation and is essential for large-scale relay systems.

Most existing rank detection methods are threshold-based [19–26], where singular values of the received signal matrix are compared with a threshold to detect the rank. The derivation of the threshold is usually based on the distribution of the noise matrix, aiming at minimizing the mean squared error (MSE) of the channel estimation [19–24] or maximizing a generalized likelihood-ratio function [25]. The limitation of existing rank detection methods is that they only consider deterministic channel matrix, where the channels are assumed to be static. Novel designs that take into account the channel randomness will further improve the rank detection accuracy.

1.2.2 Research Contributions

Aiming at fundamental studies of the promising large-scale relay network, this thesis investigates the performance analysis and transmission design for the system. The main contributions of this thesis are listed as follows.

1. **To analyze the asymptotic performance of best relay selection (BRS).** The average SNR and ergodic capacity of BRS in a single-user large-scale relay network are analyzed. Due to the complexity of the distribution of the received SNR, closed-form expressions that reveal insights for average SNR and ergodic capacity performance of BRS are not available in existing literature.

   With the help of extreme value theory (EVT) and careful manipulations of special functions, closed-form approximations for the average SNR and ergodic capacity of BRS are derived for high power range. Compared with existing results in integral forms or with special functions, our results are in closed-form and provide useful insights on the behaviour of the array gain and the ergodic capacity with respect to network parameters.

2. **To design the distributed relay beamforming scheme for multi-user relay networks and analyze its asymptotic performance.** Distributed relay beamforming is investigated in a multiple-user large-scale relay network, where users are allocated orthogonal channels to avoid interference. The

power allocation (PA) designs and performance analysis results for this relaying scheme are rare.

Firstly, the PA problem of distributed relay beamforming is solved to maximize the network throughput. Then the SNR and capacity of the proposed scheme are derived when the number of relays goes to infinity. The analytical results show that, with distributed relay beamforming in large-scale relay networks, the received SNR of each user increases linearly with the number of relay antennas and also linearly with the minimum of the user transmit power and relay power.

3. **To uncover the fundamental performance scaling-law of MRC/MRT relaying scheme with channel estimation error.** The performance of MRC/MRT is analyzed in multiple-user large-scale relay networks. Firstly, a sum-rate lower bound is derived which manifests the effect of system parameters including the number of users, the training quality, and the transmit powers of the sources and the relay. Via a general scaling model on the system parameters with respect to the number of relay antennas, the asymptotic scaling law of the network sum-rate as a function of the parameter scalings is obtained. The results show quantitatively the trade-off between the network parameters and their effects on the sum-rate. Besides, a sufficient condition for asymptotically deterministic signal-to-interference-plus-noise ratio (SINR) is shown, which covers the existing work as special cases. At last, for linearly increasing SINR, it is shown that in this case, the interference power dominates the random behaviour. Then, the outage probability and average bit error rate (ABER) are analyzed, which is not available in the literature.

4. **To design rank detection schemes for reduced-rank channel estimation.** Novel threshold-based rank detection algorithms for reduced-rank channels are proposed. Different from previous work, a system with random channel matrix model, a general training length, and unitary training matrix is considered. Lower bounds on the probability of correct rank detection are derived using the distribution of the channel matrix and noise matrix, based on which

10

the rank detection thresholds are optimized. In addition to the traditional single-threshold detection algorithm, two low-complexity multiple-threshold algorithms are further proposed. Compared with the existing schemes, our proposed schemes can achieve higher rank detection rate for various scenarios.

1.3   Organization of the Thesis

The thesis is organized as follows.

In Chapter 2, basic concepts on wireless communications, MIMO systems, and relaying schemes are introduced.

Chapter 3 represents the analysis on the average SNR and ergodic capacity of large-scale relay networks with BRS [132]. The system model, probability density function (PDF) and cumulative distribution function (CDF) analysis of the received SNR, average SNR and ergodic capacity analysis are derived. Simulations are shown to verify the analytical results.

In Chapter 4, the relay PA is investigated to maximize the network sum-rate for a multi-user multi-relay network with a total relay power constraint [121]. First, the PA problem is formulated and solved by optimization methods. Then the asymptotic performance of the designed scheme is analyzed. Simulation results are presented to show the advantage of the design and verify the analytical results.

Chapter 5 is on the performance analysis and scaling law of MRC/MRT relaying in multi-user large-scale relay networks with CSI error. Firstly, the performance scaling law is derived with all network parameters represented by the scales of the relay antenna number. Then, a sufficient condition for the SINR to be asymptotically deterministic is obtained. At last, linearly increasing SINR case is analyzed. The sufficient and necessary condition for the case is derived. Then the interference power PDF, outage probability and ABER expressions are derived and analyzed. Simulations are shown to verify the analytical results.

Chapter 6 focuses on the rank detection design for singular value decomposition (SVD)-based reduced-rank channel estimation in large-scale relay networks

[133, 134]. The channel model, training model and rank detection problem are first presented. Then a threshold selection method for single-threshold rank detection algorithm and two improved multiple-threshold rank detection methods are proposed. Finally, simulation results on the correct rank detection probability are presented to show the advantages of the proposed schemes.

Chapter 7 summarizes the main contributions of this thesis and proposes possible directions for future work.

$\sim$

Chapter 2

Background

In this chapter, we first review some basics on wireless channels and channel estimation. Then, the MIMO system and related communication models are briefly introduced. At last, important relaying schemes are described, including both distributed relaying schemes and centralized relaying schemes.

2.1 Basics of Wireless Communications

2.1.1 Characteristics of Wireless Channels

The model of wireless channel is essential for the analysis of wireless networks. Different from wired communications, signals transmitted through wireless channels may suffer from severe attenuation and distortion. The effects are characterized by two factors, i.e., large-scale fading and small-scale fading [8].

Large-Scale Fading

Large-scale fading refers to the signal attenuation caused by path loss and shadowing. Path loss measures the degradation of signals over transmission distances. Shadowing is caused by the diffraction from obstacles. The long-distance model that jointly considers both path loss and shadowing effects is given by [8]

$$P_{r(dB)} = P_{t(dB)} + P_{0(dB)} + 10 \log_{10} \left(\frac{d_0}{d} \right)^{\alpha} + X_0, \qquad (2.1)$$

where P_t is the transmit power, P_r is the received power at distance d, d_0 is the reference distance, P_0 is the path loss at the reference distance, α is the path loss

exponent, and X_0 is a zero-mean Gaussian random variable modelling shadowing effect. The value of the path loss exponent α depends on the propagation environment and usually ranges between 2 and 6.

Small-Scale Fading

Small-scale fading refers to the signal attenuation due to the presence of reflectors and scatterers that cause multiple copies of transmitted signals to arrive at the receiver. Each copy has different attenuation, delay and phase shift. Thus, these copies may add up constructively or destructively, resulting in rapid fluctuation in the signal strength.

The commonly used small-scale fading models include the Rayleigh fading model and Rician fading model. Both models assume a rich scattering environment. Thus with central limit theorem (CLT), the superposition of the channel responses from a large number of identical and independent paths are approximately Gaussian distributed.

Rayleigh fading is for the channel without a line-of-sight component. In this case, the real and imaginary part of the channel impulse response are approximately independent and identically distributed (i.i.d.) Gaussian random variables. And its phase will be evenly distributed between 0 and 2π radians. Its envelope follows the Rayleigh distribution with the PDF given by

$$f(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right), \quad x \geq 0, \tag{2.2}$$

where $\frac{\sigma^2}{2}$ is the average envelope power [8]. When represented by a complex number, the channel impulse response of Rayleigh fading is distributed as circular-symmetric complex Gaussian (CSCG), i.e., $\mathcal{CN}(0, \sigma^2)$.

Rician fading is for the channel with a dominant path, typically a line-of-sight path. In this case, the amplitude gain of the channel impulse response follows Rician distribution, whose PDF is given by [8]

$$f(x) = \frac{x}{\sigma^2} \exp\left(\frac{x^2 + A^2}{2\sigma^2}\right) I_0\left(\frac{Ax}{\sigma^2}\right), \quad x \geq 0, \tag{2.3}$$

where A is the peak amplitude of the dominant signal, and $I_0(\cdot)$ is the modified

14

Bessel function of the first kind with order 0 [27]. From the PDFs in (2.3) and (2.2) we could see that, Rayleigh fading is a special case of Rician fading with $A = 0$.

2.1.2 Channel Estimation

In most MIMO and relay communication schemes, full or partial knowledge of the channel coefficients is required at the relay, the destination, or the source for signal processing. In practice, CSI must be obtained through channel estimation.

In training-based channel estimation, known signals, which are called pilots, are sent and the channel values are estimated based on the known pilots and the received signals. In practice, due to the existence of noises, channel estimation cannot be perfect. The main metric to measure the channel estimation quality is the mean squared error (MSE). It is the average of the square of the distance between the estimation and exact parameter matrix. Denote the parameter matrix to be estimated (in MIMO training, it is the MIMO channel matrix) as $\mathbf{X}$, and the estimated matrix value (in MIMO training, it is the channel estimation) as $\hat{\mathbf{X}}$, then the MSE is formulated as follows.

$$\text{MSE} = \mathbb{E}\{\|\hat{\mathbf{X}} - \mathbf{X}\|_F^2\}. \tag{2.4}$$

In the following, two basic estimation methods, i.e., minimum-mean-squared-error (MMSE) estimator and maximal-likelihood (ML) estimator, are introduced.

MMSE and ML Estimation

MMSE estimator is the estimation that minimizes the MSE:

$$\hat{\mathbf{X}}_{MMSE} = \arg \min_{\hat{\mathbf{X}}} \text{MSE}. \tag{2.5}$$

It is proved that, the estimator is the average value of the parameters conditioned on the observation, i.e.,

$$\hat{\mathbf{X}}_{MMSE} = \mathbb{E}\{\mathbf{X}|\mathbf{Y} = \mathbf{Y}_s\}, \tag{2.6}$$

where $\mathbf{Y}$ is the observation and $\mathbf{Y}_s$ is the sample observation from $\mathbf{Y}$.

ML estimator maximizes the likelihood function. The likelihood function is defined as the probability (for discrete random variable) or probability density (for

continuous random variable) of the received symbol $\mathbf{Y}$ given the transmitted symbol $\mathbf{X}$, i.e., $f_{\mathbf{Y}}(\mathbf{Y}|\mathbf{X})$. Thus ML estimator can be represented as

$$\hat{\mathbf{X}}_{ML} = \arg\max f_{\mathbf{Y}}(\mathbf{Y}|\mathbf{X}). \tag{2.7}$$

Compared with MMSE estimator, ML estimator has lower complexity. And with ML estimator, when the number of observations goes to infinity, the estimation error is almost surely zero under mild conditions on the likelihood function [28].

2.2 Introduction of MIMO Systems

The transmission in wireless systems may fail when the channels are in deep fading. To overcome the fading effects, different diversity techniques can be exploited. For example, different versions of the same signal can be sent at different time instants to exploit the time diversity; or sent over several frequency channels to exploit the frequency diversity [29, 30]. As the time diversity and the frequency diversity are limited by the scarce time and frequency resources, to further improve the performance of wireless systems, the spatial diversity techniques are proposed, where the signal is transmitted over several different propagation paths. MIMO is the most popular spatial diversity technique, which deploys multiple antennas at the transmitter or the receiver. The research on MIMO techniques can be traced back to 1970's. It became more and more popular since early 1990's [29, 30]. Now, MIMO has become an essential element of 3G and 4G wireless communication standards [29, 30]. In what follows, the basic MIMO transmission and linear processing techniques are introduced.

2.2.1 MIMO Communications

A basic MIMO system model is shown in Fig. 2.1, where the transmitter is equipped with M antennas and the receiver is equipped with N antennas. The channels between the transmitter and the receiver can be represented by an $N \times M$ matrix $\mathbf{H}$, whose (i, j)-th entry, $h_{i,j}$, is the channel from the j-th transmitter antenna to the i-th receiver antenna. As the MIMO channel may have reduced-rank [29, 30], we assume that the rank of $\mathbf{H}$ is r ($r \leq \min(M, N)$).

16
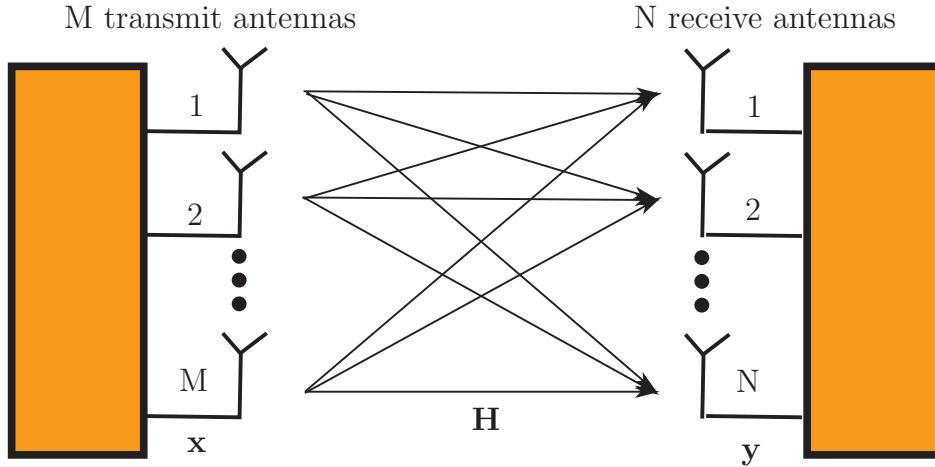
Figure 2.1: Single-user MIMO system diagram

Let $\mathbf{x}$ be the $M \times 1$ transmitted symbol vector, $\mathbf{W_t}$ be the $M \times M$ transmit pre-coding matrix, and $\mathbf{W_r}$ be $N \times N$ receive combining matrix. Then, at the receiver side, after combining, the $N \times 1$ symbol vector , $\mathbf{y}$, is

$$\mathbf{y} = \sqrt{P}\mathbf{W_r}\mathbf{H}\mathbf{W_t}\mathbf{x} + \mathbf{W_r}\mathbf{w}, \tag{2.8}$$

where P is the total transmit power and $\mathbf{w}$ is the additive noise vector, whose entries are i.i.d. Gaussian distributed with zero mean and variance σ_w^2.

With the use of channel estimation, $\mathbf{H}$ can be known at both the transmitter and receiver. We can then design the precoding and combining matrices to exploit the spatial diversity.

First, decompose $\mathbf{H}$ with SVD as

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H, \tag{2.9}$$

where $\mathbf{U}$ and $\mathbf{V}$ are $N \times N$ and $M \times M$ unitary matrices, and $\mathbf{\Sigma}$ is an $N \times M$ diagonal matrix with r non-zero singular values of $\mathbf{H}$ on its diagonal in non-increasing order.

Then, by designing the precoding and combining matrices as $\mathbf{W_t} = \mathbf{V}$ and $\mathbf{W_r} = \mathbf{U}^H$, we have

$$\mathbf{y} = \sqrt{P}\mathbf{\Sigma}\mathbf{x} + \tilde{\mathbf{w}}, \tag{2.10}$$

where $\tilde{\mathbf{w}} = \mathbf{U}^H\mathbf{w}$ is the equivalent noise vector. It can be shown that entries of $\tilde{\mathbf{w}}$ are uncorrelated. Since $\mathbf{\Sigma}$ is a diagonal matrix with r non-zero elements, we

17

have produced equivalently r parallel channels with uncorrelated noises. So r is the maximum number of symbols can be transmitted at one time such that the receiver is able to successfully separate all transmitted symbols without interference.

With r parallel sub-channels available, the MIMO system can be designed to achieve the spatial diversity gain or spatial multiplexing gain as described below.

- Spatial diversity gain is achieved by transmitting different versions of the same symbol over multiple sub-channels. With appropriate combining at the receiver, the transmission reliability will be improved, resulting in lower symbol error probability. The maximum possible spatial diversity order is shown to be the product of the number of transmit and receive antennas (i.e. MN). It is achievable when all channel coefficients are independently distributed.

- Spatial multiplexing gain is achieved by transmitting multiple symbols over the sub-channels. This results in throughput improvement without extra cost of power or frequency bandwidth. The maximum spatial multiplexing gain achieved by the MIMO system described above is r.

Obviously, the maximum spatial diversity gain and spatial multiplexing gain cannot be achieved simultaneously. A natural trade-off exists between the two benefits [29, 30].

2.2.2 Performance Metrics

To evaluate the performance of a MIMO or relay system, performance metrics have been proposed. The popular ones include the average SNR, ergodic capacity, outage probability, ABER, correct rank detection probability and asymptotic parameter scaling.

- The average SNR is the expected value of the SNR over all channel states. It shows the average communication quality. Besides, from the average SNR, we can find the array gain for MIMO or virtual MIMO systems. Array gain is defined as the power gain of using multiple antennas or multiple virtual communication paths over single-antenna systems. It can be obtained by dividing the average SNR with the average SNR of single-antenna system.

18

- The ergodic capacity is the average Shannon capacity over all channel states. Shannon capacity is defined as the maximum data rate of error-free transmission with specified bandwidth. For fading channels with additive white Gaussian noise, the ergodic capacity, denoted as C_{erg}, is given by

$$C_{erg} = B \int_0^\infty \log_2(1 + \gamma) f(\gamma) d\gamma, \qquad (2.11)$$

where B is the bandwidth in hertz, γ is the received SNR, and $f(\gamma)$ is the PDF of the received SNR.

Besides, from the ergodic capacity, the multiplexing gain can be calculated. With the ergodic capacity represented as a function of the average transmit SNR, i.e., $C_{erg}(\text{SNR}_t)$, the multiplexing gain is defined by [29]

$$g_m = \lim_{\text{SNR}_t \to \infty} \frac{C_{erg}(\text{SNR}_t)}{\log_2 \text{SNR}_t}. \qquad (2.12)$$

- The outage probability is defined as the probability that the instantaneous SNR falls below a certain threshold. The threshold is often selected to guarantee the desired QoS. Thus, the outage probability is an important measure for the quality of communication systems. From the definition, it can be found directly from the CDF of the SNR:

$$P_{out}(\gamma_{th}) = P(\text{SNR} < \gamma_{th}), \qquad (2.13)$$

where γ_{th} is the predefined threshold and $P(\cdot)$ is the CDF of the SNR.

From the outage probability, the diversity gain can be obtained as below, when the outage probability is represented as a function of the transmit SNR.

$$g_d = - \lim_{\text{SNR}_t \to \infty} \frac{\log_2 P_{out}(\gamma_{th}, \text{SNR}_t)}{\log_2 \text{SNR}_t}. \qquad (2.14)$$

- The ABER, also known as bit error probability, is the probability that a bit is incorrectly decoded at the receiver. It depends on the SNR at the receiver as well as the modulation and coding schemes. To derive the average BER, we need to average the BER over the distribution of the SNR, i.e.,

$$P_b = \int_0^\infty P_b(e|\gamma) f(\gamma) d\gamma,$$

19

where $P_b(e|\gamma)$ is the BER conditioned on the instantaneous SNR.

With ABER represented as a function of the transmit SNR, we have an alternative definition of the diversity gain as

$$g_d = - \lim_{\text{SNR}_t \to \infty} \frac{\ln P_b(\text{SNR}_t)}{\ln \text{SNR}_t}. \tag{2.15}$$

- Correct rank detection probability is the probability that the channel rank is correctly detected. For reduced-rank MIMO channel, rank detection is an important factor of channel estimation. It is also essential for transmission design as channel rank indicates the maximum number of symbols can be reliable communicated within the same time-frequency bandwidth.

- Asymptotic parameter scaling shows how the network performance metrics scale with network parameters when the parameters approach certain values asymptotically. It is important for the performance analysis of MIMO systems, especially massive MIMO systems and will help guide the design of such networks.

2.3   Relaying Schemes

As introduced in the previous section, MIMO systems exploit the spatial diversity by embedding multiple antennas on the transceivers. However, for many applications, e.g., in sensor networks or for personal mobile devices, due to the limited size and cost, it may not be practical to place multiple antennas on the transceiver devices. In this case, an alternative technology to achieve spatial diversity is to use relays, where the relay antennas can form a virtual MIMO antenna array. In this section, relay networks and important relaying schemes will be introduced.

As has been shown in Fig.1.3, there are different types of relay networks. Basically, they can be classified as the infrastructure-based relay network and the cooperative relay network, as shown in Fig.2.2. In the figure, 'BS' stands for base station, and 'RS' stands for relay station. In infrastructure-based relay network, a supportive relay station is deployed in the network to help the communication between the users with the base station (Fig.2.2a) or the communication between
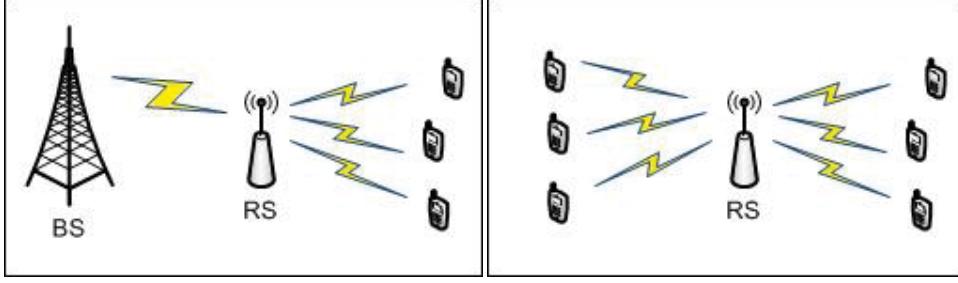
(a) Infrastructure-based relay network with a base station
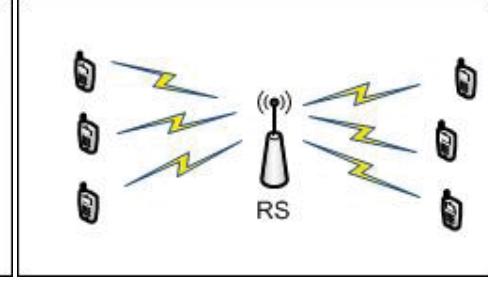
(b) Infrastructure-based relay network for D2D communication
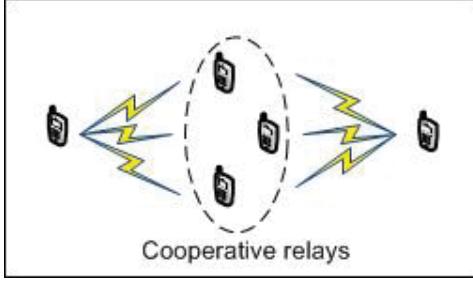
(c) Cooperative relay network

Figure 2.2: Different types of relay networks.

users directly (Fig.2.2b). Unlike the base station, the relay station usually has a smaller size, less transmit power, and wireless backhaul connections. Thus it is less expensive and much easier to deploy in the network [8]. Different from the infrastructure-based relay network, the cooperative relay network does not have a deployed relay station, and instead, the user devices act as relay nodes to help transmit each other's signal. Fig.2.2c shows that a group of users are helping relaying signals from a source node to a destination node. Compared with the infrastructure-based relay network, the cooperative relay network is more flexible and costs less to implement. But due to the opportunistic and dynamic nature of the relay nodes, the cooperative relay network is not as reliable as the infrastructure-based relay network. In addition, as most MIMO technologies can be directly applied at the relay station in the infrastructure-based relay network, thus this network is also referred to as the MIMO relay network. When the number of relay antennas is very large, it is called the massive MIMO relay network.

In both kinds of relay networks, the transmission of signals from the sources to the destinations takes two steps. In the first step, the sources transmit signals to the

relays. In the second step, the relays process their received signals, and transmit the processed signals to the destinations.

The processing at relays, referred to as relaying schemes, is a key component of relay technology. The original studies of relay date back to 1970's [31, 32], where properties of relay channels are investigated. The relay technology becomes very popular in 2000's, and various protocols and schemes have been proposed for relay communications [11, 33–35]. Two basic ones are amplify-and-forward (AF) and decode-and-forward (DF).

- **Amplify-and-Forward**. In AF, the relays amplify the received signals from the sources and forward them to the destinations, without explicitly decoding or demodulating the messages or symbols. It is often referred to as non-regenerative relaying schemes. Compared with DF, AF requires simpler processing, which makes for low-cost implementation and short processing delay for relaying [8]. Its disadvantage is that besides the desired signals, noise and interference are also amplified by the relays [8], causing degradation in the performance.

- **Decode-and-Forward**. In DF, the relays first decode the received signals then retransmit the decoded information to the destinations with or without re-encoding. Compared with AF, DF overcomes the drawback of noise and interference amplification, thus can result in better throughput [8]. But DF relaying is more complex than AF and prone to error propagation and processing delays [8].

Compared with DF, AF is preferred for large-scale relay networks for its simplicity and low cost, as well as high asymptotic performance [36–45]. Thus, in this thesis, we mainly focus on AF relaying schemes, where the relay processes and retransmits the signals without decoding.

Next, a few popular relaying schemes will be introduced, including distributed relaying schemes and centralized relaying schemes.

2.3.1 Distributed Relaying Schemes

Distributed relaying schemes are mainly designed for the cooperative relay networks (Fig. 2.2c). In a distributed relaying scheme, each relay node only has its own channel information, and there is no or limited information sharing between the relays. This feature reduces the overhead of inter-communications between relays, making it possible to be implemented in networks with a large number of relays. Besides, each relay only performs simple signal processing, e.g., angle adjustment and power amplification, making it fit for devices with limited power and processing capability. Popular distributed relaying schemes include orthogonal cooperation, distributed space-time coding, relay selection and distributed relay beamforming [8]. In this thesis, we mainly consider relay selection and distributed relay beamforming. The orthogonal cooperation is not considered as it requires all sources and relays transmit over orthogonal time or frequency channels [8], which is impractical for large-scale relay networks when the number of relay antennas is large. While distributed space-time coding can achieve full diversity gain without knowledge of the instantaneous CSI at the relay [8], its high complexity and involved code design for high dimension makes it inappropriate for large-scale relay networks. In the following, the two most-commonly used relaying schemes, relay selection and distributed relay beamforming, will be introduced.

Relay Selection

In relay selection, a subset of relays are chosen to cooperate. The most prominent advantages of relay selection are its low overhead and low synchronization requirements. It has also been shown that proper selection of relays can lead to high performance such as full diversity [11–14]. Compared with schemes where the relays transmit in orthogonal channels, relay selection can achieve higher spectral efficiency [33]. Thus, relay selection is a highly practical relaying scheme with good potentials in performance.

The most commonly-used relay selection scheme is BRS, where the best relay path with the highest SNR is selected. In [11], with rigorous derivations on error probability, BRS is proved to achieve full diversity order. The outage probability,

average SNR, ergodic capacity, and symbol error probability are further analysed in [15–18]. In [46], from the information theory point of view, the capacity scaling law of BRS with respect to the number of relays are shown for large-scale relay networks.

Distributed Relay Beamforming

Different from relay selection, in distributed relay beamforming, all relays participate in the communications using the same frequency bandwidth. Although each relay only has single antenna, the cooperation between relays can form a virtual antenna array. Thus beamforming can be applied distributively at the relays.

In distributed relay beamforming, each relay adjusts the power and phase of its received signal according to its own channel information [34, 35]. With perfect phase alignment, distributed relay beamforming usually reduces to PA. In [35] and [34], for single-antenna multiple-relay networks with perfect CSI, closed-form solutions were provided for the relay beamforming vector under a total relay power constraint and per-relay power constraints, respectively. [12, 47, 48] investigated distributed beamforming with partial CSI. Relay beamforming and PA for multiple-user multiple-relay networks were studied in [49–58].

2.3.2 Centralized Relaying Schemes

Centralized relaying schemes are mainly proposed for infrastructure-based relay networks (Fig. 2.2a, Fig. 2.2b), where a multiple-antenna relay station is installed. It can be used to assist communications of cell-edge users, mobile users, or D2D terminals. Unlike distributed relaying schemes, the overhead of inter-communications of relay antennas is not a concern in centralized relaying schemes, because all relay antennas are connected to a central controller. In centralized relaying schemes, the central controller first combines the received signals from all relay antennas using the CSI of source-to-relay channels. Then the combined signals are precoded based on the CSI of relay-to-destination channels and sent to the destinations. Compared with distributed relaying schemes, where each relay antenna only processes its local received signal with local CSI, centralized relaying schemes can better combine and

precode the signals to further reduce the effect of small-scale fading and inter-user interference. For networks with a centralized multi-antenna relay, the source-to-relay transmission alone is equivalent to the uplink MIMO transmission, and the relay-to-destination transmission alone is equivalent to the downlink MIMO transmission. But different from a simple decoupled combination of MIMO uplink and downlink transmission, in the second step, the interference and noise in the first step is also amplified, making its performance more complex to analyse. But, due to the similarity to MIMO transmissions, typical MIMO uplink precoding and downlink receiving structures or concepts can be used for the relay network. Both linear processing schemes and non-linear processing schemes have been proposed for MIMO transmissions. The non-linear processing schemes, e.g., ML detection [59] and dirty paper coding [60, 61], can usually better mitigate inter-user interference and achieve higher performance than the linear ones, but at the cost of considerable higher complexity. Besides, in large-scale networks, the linear processing schemes can achieve comparable performance to the non-linear processing schemes [36]. Thus the linear processing schemes are more preferred. Popular linear processing schemes include MRC/MRT and zero-forcing (ZF) processing schemes. In the following, the two schemes will be introduced.

- **MRC/MRT processing.** In MRC/MRT, the signals received from all relay antennas are firstly combined following the MRC rule. The MRC rule coherently combines the signals from the same source but received by different antennas. This is done by multiplying the received signal vector with the MRC matrix, which is the conjugate transpose of the source-to-relay channel matrix. Then, the combined signals are precoded following the MRT rule. The MRT rule is designed for the intended signals received from different wireless links to be added together coherently at the target destination. This is done by MRT precoding, where the combined signal vector is multiplied with the conjugate transpose of the relay-to-destination channel matrix before transmitted to the destinations. With the source-to-relay channel denoted as $\mathbf{F}$, and the relay-to-destination channel denoted as $\mathbf{G}$, the signal processing

matrix for MRC/MRT is

$$\mathbf{W}_{\mathrm{mrcmrt}} = \mathbf{G}^H \mathbf{F}^H. \tag{2.16}$$

- **Zero-forcing processing.** The principle of ZF is to nullify multi-user inter-ference by using the CSI. In ZF relaying, firstly, the signals received at the re-lay antennas are combined by multiplying the received signal vector with the pseudo inverse of source-to-relay channel matrix. Then, the combined signals are precoded by multiplying with the pseudo inverse of relay-to-destination channel matrix. The signal processing matrix for ZF is

$$\mathbf{W}_{\mathrm{zf}} = \mathbf{G}^H \left(\mathbf{G}\mathbf{G}^H \right)^{-1} \left(\mathbf{F}^H \mathbf{F} \right)^{-1} \mathbf{F}^H. \tag{2.17}$$

Generally, ZF relaying outperforms MRC/MRT relaying due to its ability to to-tally cancel inter-user interference. While, MRC/MRT relaying is better than ZF in its robustness and less computation burden. In large-scale relay networks, both schemes can achieve high energy and spectral efficiency [36]. While, ZF outper-forms MRC/MRT for high spectral efficiency and low energy efficiency scenarios, but in low spectral efficiency and high energy efficiency scenarios, the opposite holds [62].

$$\sim$$

Chapter 3

Closed-Form Average SNR and Ergodic Capacity Expressions for Best Relay Selection

In this chapter, we analyse the average SNR and ergodic capacity of large-scale relay networks with BRS. First, EVT is used to obtain an implicit expression for the asymptotic CDF of the received SNR when the number of relays is large. Then via high power approximations, closed-form expressions for the CDF and PDF of the received SNR are achieved, from which closed-form expressions of the average received SNR and ergodic capacity are derived. Insights on the array gain and ergodic capacity behaviour of BRS can be observed from the derived results. Simulations show that the derived approximations are tight, even for not-so-large relay networks.

3.1 Introduction

As introduced in Section 2.3.1, BRS is one of the most popular relaying schemes for its low overhead and high performance. It has been widely investigated in the literature [11, 13, 15–18, 46]. In [11], with rigorous derivations on error probability, BRS was proved to achieve full diversity order. In [15], exact expressions of the outage probability, average SNR, ergodic capacity, and symbol error probability were derived. However, the results are represented as summations of infinite series, with special functions, or in integral forms. They are not in closed-form and provide

little insights. Other work on performance analysis uses approximations. In [17,18], the received SNR of a relay path is approximated as the smaller value of the SNRs of the source-to-relay and relay-to-destination links. Closed-form approximations on the outage probability, symbol error probability, average SNR, and ergodic capacity were derived for Rayleigh fading channels. Although the approximations on the error rate and outage probability are tight, the average SNR and ergodic capacity approximations are loose since the derived CDF and PDF of the received SNR are only tight around the origin. Another approximation is by ignoring a noise term in the received SNR [16]. But results from this approximation are in forms of products/summations of special functions and little insights can be obtained. The only work gives insight on the ergodic capacity is [46]. From the information theory point of view, [46] showed the capacity scaling law with respect to the number of relays for large-scale relay networks. However, closed-form expression for the ergodic capacity was not available.

Thus, while the symbol error rate and outage probability of BRS for high power region have been well-understood (tight closed-form approximations were derived), existing closed-form expressions for the average SNR and ergodic capacity are either not tight [17, 18] or in forms of products/summations of special functions [16], which give little insights. The main reason is the complexity of the distribution of the received SNR for relay networks. For symbol error rate or outage probability, it was proved in [14] that they are dominated by the behaviour of the SNR distribution around zero only, which can be derived in tractable form. For average SNR or capacity, however, the behaviour of the SNR distribution in the medium or large value range is more important, which is difficult to derive or approximate.

In this chapter, with the help of EVT, we derive tight and closed-form approximations on the average SNR and ergodic capacity of BRS in large-scale relay networks [35, 36, 63–69]. EVT studies the behaviour of the maxima of i.i.d. random variables [71]. It has been used to analyse the performance of broadcast MIMO networks [72] and spectrum sharing multi-hop relay networks [73]. We first obtain an asymptotic and implicit expression for the CDF of the received SNR via EVT. Then with high power approximations, we derive tractable and closed-form
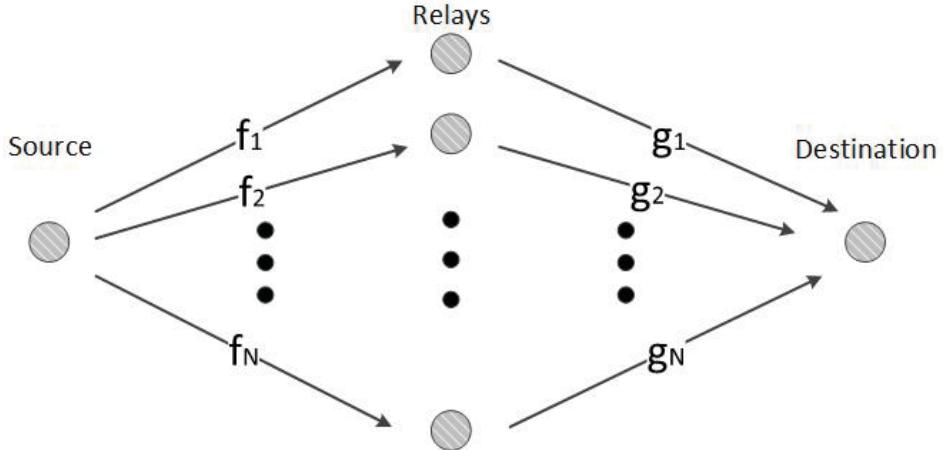
Figure 3.1: System model for single-user multiple-relay networks

approximations of the CDF and PDF of the received SNR. Simulation shows that the approximations are tight for the whole variable range (not just around zero). Using the same approximations, closed-form expressions on the average SNR and ergodic capacity are derived, from which their behaviour with respect to network parameters such as the number of relays and transmit power can be observed. Simulation shows that our derived approximations can well represent the performance of BRS, even for not-so-large relay networks.

The rest of the chapter is organized as follows. In Section 3.2, the relay network model and BRS are introduced. Section 3.3 is on the derivations of closed-form approximations on the PDF and CDF of the received SNR. Section 3.4 contains closed-form approximations of the average SNR and ergodic capacity. In Section 3.5, simulations are shown. In Section 3.6, we summarize this chapter.

3.2 System Model

We consider a network model as shown in Fig. 3.1, with one single-antenna source, one single-antenna destination, and N single-antenna relays ($N \to \infty$ for large-scale relay networks). Denote the channel from the source to the nth relay as f_n and the channel from the nth relay to the destination as g_n. All channels are assumed to be independent and follow zero-mean CSCG distribution, i.e., Rayleigh fading channels. We assume that the distance from the source/destination to differ-

ent relays are about the same. Thus, f_i's have the same variance, denoted as σ_f^2; g_i's have the same variance, denoted as σ_g^2. One application is for networks where the source and destination are far apart and a cluster of relays (close to each other) are available to help their communications. Denote the transmit power of the transmitter as P and the transmit power of each relay as Q. All noises are assumed to be i.i.d. and follow $\mathcal{CN}(0, 1)$.

With the two-step BRS scheme described in Chapter 2, the end-to-end received SNR via the nth relay path, denoted as SNR_n, is well known to be

$$\mathrm{SNR}_n = \frac{PQ|f_n g_n|^2}{1 + P|f_n|^2 + Q|g_n|^2}, \tag{3.1}$$

Note that SNR_n's are i.i.d. due to the independence of different relay paths. In BRS, the relay path with the maximum end-to-end received SNR is selected. The received SNR after selection, denoted as SNR_S, is

$$\mathrm{SNR}_S = \max_{n=1,2,\ldots,N} \{\mathrm{SNR}_n\}. \tag{3.2}$$

3.3 PDF and CDF Analysis of the Received SNR

In this section, we derive closed-form approximations on the CDF and PDF of SNR_S using EVT and high power approximations.

We first review the EVT results needed for this work, which are from Theorem 3.1 and Proposition 3.3 of [70] and related results in [71]. Let $X_1, X_2, \ldots, X_N$ be a sequence of N i.i.d. random variables whose CDF and PDF, denoted as F_X and f_X respectively, satisfy

$$\lim_{x \to \infty} \frac{f_X'(x)[1 - F_X(x)]}{f_X^2(x)} = -1. \tag{3.3}$$

Let $Y \triangleq \max\{X_1, X_2, \ldots, X_N\}$. When $N \to \infty$, the CDF of Y can be derived as follows

$$\lim_{N \to \infty} F_Y(y) = \exp\left(-\exp\left[-\frac{y - b_N}{a_N}\right]\right),$$

where

$$b_N = F_X^{-1}(1 - 1/N), \; a_N = \frac{1 - F_X(b_N)}{f_X(b_N)} \tag{3.4}$$

30

with $F_X^{-1}(\cdot)$ the inverse function of $F_X(\cdot)$.

With the EVT results, we can derive the CDF of SNR_S, which is the maxima of SNR_n's We first verify that the CDF and PDF of SNR_n satisfy (3.3). The exact CDF and PDF of SNR_n are as the following [12, 15].

$$F(x)=1-\exp\left(-\frac{x(1+\eta)}{\eta\sigma_f^2P}\right)\frac{2\sqrt{x(1+x)}}{\sqrt{\eta}\sigma_f^2P}K_1\left(\frac{2\sqrt{x(1+x)}}{\sqrt{\eta}\sigma_f^2P}\right),\qquad(3.5)$$

$$f(x)=\frac{2\exp\left(\frac{-x(1+\eta)}{\eta\sigma_f^2P}\right)}{\eta^{3/2}\sigma_f^4P^2}\left[(1+\eta)\sqrt{x(1+x)}K_1\left(\frac{2\sqrt{x(1+x)}}{\sqrt{\eta}\sigma_f^2P}\right)\right.$$
$$\left.+\sqrt{\eta}(1+2x)K_0\left(\frac{2\sqrt{x(1+x)}}{\sqrt{\eta}\sigma_f^2P}\right)\right],\qquad(3.6)$$

where η is defined as

$$\eta\triangleq\frac{\sigma_g^2}{\sigma_f^2}\cdot\frac{Q}{P}.\qquad(3.7)$$

Lemma 1. *The exact CDF and PDF of* SNR_n *satisfy the condition in (3.3).*

Proof. The proof is available in Appendix A. ☐

So, when $N\to\infty$, the CDF of SNR_S is

$$\lim_{N\to\infty}F_{\mathrm{SNR},S}(x)=\exp\left(-\exp\left[-\frac{x-b_N}{a_N}\right]\right),\qquad(3.8)$$

where a_N and b_N are given in (3.4) with F_X and f_X replaced by the exact CDF and PDF of SNR_n given in (3.5) and (3.6).

The right-hand-side of (3.8) is the CDF of a Gumbel distribution whose scale parameter is a_N and location parameter is b_N. For finite N, the CDF in the right hand side of (3.8) is an approximation.

The asymptotic CDF result in (3.8) is obtained by direct use of EVT after verifying the condition in (3.3). Due to the modified Bessel function in $F(x)$ and $f(x)$, as well as the inverse function in b_N, the CDF in (3.8) is in implicit form, not explicit closed-form. Its use in further analysis of the average SNR and ergodic capacity cannot lead to insightful results. Thus, in what follows, we derive tight approximations on a_N, b_N to obtain closed-form CDF and PDF formulas. This is also our

31

main technical contribution in this Section. The results are presented in Proposition 1.

Proposition 1. *Define*

$$C_1 \triangleq \frac{\eta^{\frac{1}{4}}}{1 + \sqrt{\eta}} \tag{3.9}$$

and

$$C_2 \triangleq 2.9753 \left[1 - \left(1 + \frac{0.3361\sqrt{\ln(NC_1) - 0.2742}}{1 - 0.0042\left[2\ln(NC_1) - 0.5484\right]\exp\left[-0.0201\sqrt{2\ln(NC_1) - 0.5484}\right]} \right)^{-1} \right]. \tag{3.10}$$

When $N \to \infty$, $P, Q \gg 1$, and η is bounded, the CDF of SNR_S can be approximated as

$$F_{\mathrm{SNR},S}(x) \approx \exp\left[-\sqrt{\frac{\pi}{2}} N C_1 \exp\left(C_2 - \frac{1}{\sigma_f^2 \sqrt{\eta} C_1^2} \frac{x}{P} \right) \right]; \tag{3.11}$$

the PDF of SNR_S can be approximated as

$$f_{\mathrm{SNR},S}(x) \approx \sqrt{\frac{\pi}{2}} \frac{N}{\sigma_f^2 \sqrt{\eta} C_1 P} \exp\left[-\sqrt{\frac{\pi}{2}} N C_1 \exp\left(C_2 - \frac{x}{\sigma_f^2 \sqrt{\eta} C_1^2 P} \right) - \frac{x}{\sigma_f^2 \sqrt{\eta} C_1^2 P} + C_2 \right]. \tag{3.12}$$

Proof. The proof is available in Appendix B. $\qquad\qquad\square$

Our main effort in the proof of Proposition 1 is to find tight closed-form approximations of a_N and b_N. There are other ways for their approximations, for example, by annealing simulated PDF (histogram) of a given network with the CDF structure in (3.8), or numerically finding a_N and b_N by solving Equations (B.1) and (3.4). However, with these methods, the approximations can only be done trial by trial. A new trial is needed for a change in the network setting (e.g., channel covariances, source/relay power, or network size). Besides, analytical formulas for the PDF and CDF cannot be obtained. On the contrast, our results in Proposition 1 are in analytical closed-form and allow further derivations on the average SNR and ergodic capacity.

In the literature, for performance analysis of relay networks, two widely used approximations on SNR_n are used. One is the harmonic mean [16], where SNR_n in (3.1) is approximated with the following upper bound:

$$\text{SNR}_n \leq \frac{PQ|f_n g_n|^2}{P|f_n|^2 + Q|g_n|^2}.$$

The other is the minimum SNR of the source-relay and relay-destination links [17, 18], which is also an upper bound of SNR_n:

$$\text{SNR}_n \leq \min\{P|f_n|^2, Q|g_n|^2\}.$$

The derived CDF/PDF approximations corresponding to the two approximations are, respectively,

$$F_{\text{SNR},S,harmonic} \approx \left[1 - \exp\left(-\frac{x(1+\eta)}{\eta\sigma_f^2 P}\right)\frac{2x}{\sqrt{\eta}\sigma_f^2 P}K_1\left(\frac{2x}{\sqrt{\eta}\sigma_f^2 P}\right)\right]^N, \quad (3.13)$$

$$f_{\text{SNR},S,harmonic} \approx \frac{2Nx}{\eta\sigma_f^4 P^2}\exp\left(-\frac{1+\eta}{\eta\sigma_f^2 P}x\right)\left[\frac{1+\eta}{\sqrt{\eta}}K_1\left(\frac{2x}{\sqrt{\eta}\sigma_f^2 P}\right)+2K_0\left(\frac{2x}{\sqrt{\eta}\sigma_f^2 P}\right)\right] \times$$

$$\left[1 - \exp\left(-\frac{x(1+\eta)}{\eta\sigma_f^2 P}\right)\frac{2x}{\sqrt{\eta}\sigma_f^2 P}K_1\left(\frac{2x}{\sqrt{\eta}\sigma_f^2 P}\right)\right]^{N-1}, \quad (3.14)$$

and

$$F_{\text{SNR},S,min}(x) \approx \left[1 - \exp\left(-\frac{1+\eta}{\eta\sigma_f^2 P}x\right)\right]^N, \quad (3.15)$$

$$f_{\text{SNR},S,min}(x) \approx \frac{1+\eta}{\eta\sigma_f^2 P}N\exp\left(-\frac{1+\eta}{\eta\sigma_f^2 P}x\right)\left[1 - \exp\left(-\frac{1+\eta}{\eta\sigma_f^2 P}x\right)\right]^{N-1}. \quad (3.16)$$

Compared with the exact PDF/CDF formulas in [15] and the harmonic mean approximations, both containing products of N modified Bessel functions, our results are more tractable and favourable in further analysis of the average SNR and ergodic capacity. The minimum link-SNR approximations have simpler expressions than our results. But they are tight for x close to 0 only. For other values of x, the approximations are loose. Our CDF and PDF results are tight for all values of the received SNR, when the transmit power is high. It is noteworthy that since the channels are random variables, the high transmit power condition and the high SNR

condition are different. Even with high transmit power, the received SNR value may be small when channels are in deep fading.

Our derived approximations converge to the exact ones only when the number of relays grows to infinity. Thus it can be used to analyze the asymptotic performance of the large-scale relay networks [35, 36, 63–69]. But simulation shows that they are also tight for a wide practical range of relay numbers such as $N = 10$ or 20. So they can be useful for not-so-large relay networks as well.

3.4 Average Received SNR and Ergodic Capacity Analysis

In this section, we analyse the average received SNR and ergodic capacity using the closed-form CDF and PDF approximations in Proposition 1, and discuss their behaviour.

For the Gumbel CDF in (3.8) with scale parameter a_N and location parameter b_N, its average is [75]

$$\mathbb{E}(\text{SNR}_S) = b_N + \gamma a_N,$$

where γ is the Euler constant. Thus, with straightforward calculations, when $N \to \infty$, $P, Q \gg 1$, and η is bounded, we can use the approximations of a_N and b_N in (B.8) and (B.7) to obtain the following closed-form approximation on the average received SNR of BRS:

$$\mathbb{E}\{\text{SNR}_S\} \approx \sqrt{\eta} C_1^2 \sigma_f^2 \left[\ln N + \ln C_1 + C_2 + \frac{1}{2} \ln \frac{\pi}{2} + \gamma \right] P, \qquad (3.17)$$

where C_1 and C_2 are defined in (3.9) and (3.10).

While existing results on the average received SNR based on the exact SNR distribution and its harmonic mean approximation are in integral forms, or in summation forms with special functions [15, 16], our result is in closed-form. If the minimum link-SNR approximation is used, we can obtain the following neat formula for the average SNR [18]:

$$\mathbb{E}\{\text{SNR}_{S,min}\} \approx \left(\sum_{n=1}^{N} \frac{1}{n} \right) \frac{\eta \sigma_f^2}{1 + \eta} P. \qquad (3.18)$$

But this approximation has considerable difference to the simulated average SNR, especially when the transmit power is high. Compared to it, our results are significantly tighter for a wide range of network scenarios, as will be shown by the simulation results in Section 4.5.

Now we discuss the behaviour of the average SNR with respect to the number of relays N and transmit power P using (3.17). First, from the definition in (3.9), C_1 is independent of N, but only a function of η defined in (3.9). Although C_2 depends on N, we can show easily that $C_2 \in (0, 2.9753]$. Thus for large N, the average SNR is of the order of $\ln N$. Also, we can see that with fixed η, the average SNR is linear in P. Our results can help the designs of relay networks. For example, we can use the formula to get the required number of relays and transmit power for given SNR requirement, or to decide whether BRS can meet service requirement.

Next, we compare the array gains [29] of BRS and a few other systems to understand the quantitative improvement/impairment in the received SNR with respect to several system settings. To focus on the difference in transmission schemes and systems, we consider the homogeneous case where $\sigma_f^2 = \sigma_g^2 = \sigma^2$ and $P = Q$. Thus $\eta = 1$. In all systems, the channels are assumed to be i.i.d. Rayleigh fading following $\mathcal{CN}(0, \sigma^2)$.

1. *Array gain of BRS in an N-relay network*: From (3.17), the array gain of BRS is

$$G_{BRS} = \frac{\sigma^2}{4}\left[\ln N + \ln\frac{1}{2} + C_2 + \frac{1}{2}\ln\frac{\pi}{2} + \gamma\right]. \qquad (3.19)$$

   To our best knowledge, closed-form result for the array gain of BRS was unavailable in existing literature. We can see that it is linear in $\ln N$ when N is large.

2. *Array gain of single-relay AF network*: For a single-relay AF network, where both the source power and the relay power are P, the array gain is $G_{SR} = \sigma^2/3$ [76]. From the comparison with BRS we know that, although in both schemes only one relay participate, the array gain improvement of BRS is

$$\frac{G_{BRS}}{G_{SR}} = \frac{3}{4}\left[\ln N + \ln\frac{1}{2} + C_2 + \frac{1}{2}\ln\frac{\pi}{2} + \gamma\right] \approx \frac{3}{4}\ln N$$

35

for large N. This improvement is due to the spatial diversity provided by the N relays.

3. *Array gain of distributed relay beamforming* [34, 35]: In distributed relay beamforming, all relays cooperate with perfect synchronization and phase and power coordination. For an N-relay network, where both the source power and the total relay power are P, the array gain of distributed relay beamforming is $G_{DRBF} = N\sigma^2/3$ [76]. Compared with distributed relay beamforming, the impairment in the array gain of BRS is

$$\frac{G_{DRBF}}{G_{BRS}} \approx \frac{4N}{3 \ln N}.$$

This is due to different channel information and synchronization requirements. For distributed relay beamforming, each relay needs to know its channels and perfectly synchronize with others; while for BRS, no channel information or synchronization is needed at the relays.

4. *Array gain of MISO system with antenna selection*: For an $N \times 1$ MISO system, antenna selection chooses the transmit antenna leading to the highest received SNR. The array gain is known as $G_{MISO-AS} = \sigma^2 \sum_{n=1}^{N} \frac{1}{n} \approx \sigma^2 \ln N$. Compared with the MISO system with antenna selection, the impairment in the array gain of BRS in a relay network is

$$\frac{G_{MISO-AS}}{G_{BRS}} \approx 4 \approx 6\text{dB}.$$

for large N. This impairment is due to the relay noise amplification in a relay network.

While the array gain comparison of single-antenna case, antenna selection, and beamforming is available for multiple-antenna systems [29], it is unavailable for relay networks due to the complexity of the SNR distribution. Our results fill this missing part.

Now, we analyse the ergodic capacity of BRS. When $N \to \infty$, $P, Q \gg 1$, and η is bounded, from (3.17) and Jensen's inequality, we obtain the following

36

approximation:

$$C_{Erg} \leq \frac{1}{2} \log_2 \left(1 + \mathbb{E}\{\text{SNR}_S\} \right)$$
$$\approx \frac{1}{2} \log_2 \left[1 + \sqrt{\eta} C_1^2 \sigma_f^2 \left(\ln N + \ln C_1 + C_2 + \frac{1}{2} \ln \frac{\pi}{2} + \gamma \right) P \right], \qquad (3.20)$$

where C_1 and C_2 are defined in (3.9) and (3.10).

We can see from (3.20) that the ergodic capacity scales as $\log P$ for large P and $\log \log N$ for large N. This coincides with the capacity scaling law in [46], which shows that when $N \gg 1$ and P fixed, the capacity of BRS scales as $\frac{1}{2} \log \log N + \mathcal{O}(1)$. Compared with this result, our approximation provides more details of the capacity and is a tight approximation for a wider parameter range. Our result can help the designs of networks, e.g., to achieve a certain ergodic capacity, how many relays to deploy and how much power to use.

If the minimum link-SNR approximation is used, the following approximation can be obtained similarly:

$$C_{Erg,min} \approx \frac{1}{2} \log_2 \left[1 + \left(\sum_{n=1}^{N} \frac{1}{n} \right) \frac{\eta \sigma_f^2}{1 + \eta} P \right]. \qquad (3.21)$$

This result shows the same asymptotic behaviour of the capacity with respect to P and N. But it is a very loose approximation of the capacity, while our result in (3.20) is significantly tighter and can approximate the capacity values for a wide range of network scenarios.

3.5   Simulation Results

In this section, we show simulation results to verify the derived approximations and compare with existing results.

In Fig. 3.2 and Fig. 3.3, we show the closed-form CDF and PDF approximations of the received SNR in (3.11) and (3.12) and compare with three cases: 1) the simulated CDF and PDF; 2) the CDF and PDF in [17]; 3) the CDF and PDF in [16]. In both figures, the number of relays is set as 20. The transmit powers of the source and relays are set to be the same. We use 10dB and 15dB in the simulation. From the figures, we see that our closed-form approximations match the simulated ones

37
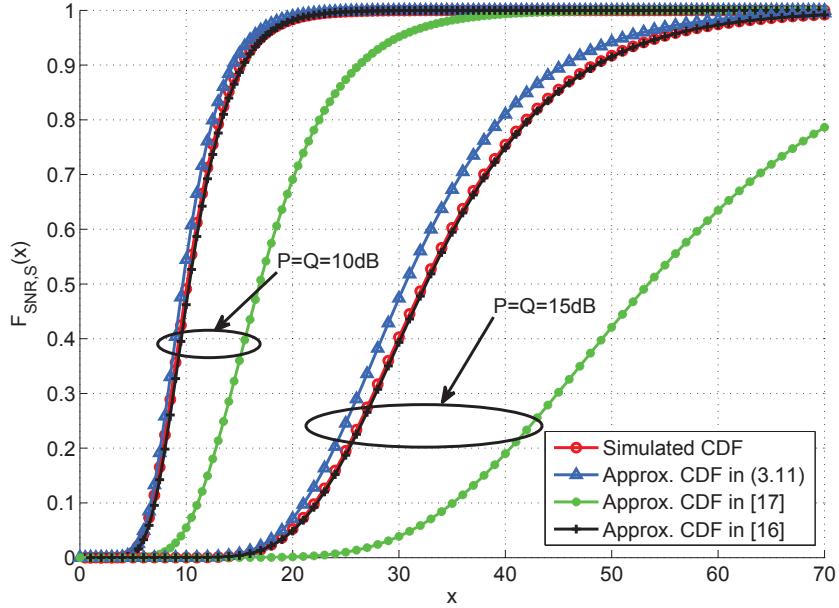
Figure 3.2: CDF of the received SNR for networks with 20 relays.

well for all values of x. The CDF and PDF approximations in [16] are tighter than the proposed ones. However, they are less tractable and not favourable for further analysis. The CDF and PDF approximations in [17] are only tight for x around zero, which is why they can be useful in the outage probability and error rate analysis, but do not work well for the average SNR and ergodic capacity analysis.

In Fig. 3.4 and Fig. 3.5, we show the average received SNR derived in (3.17) and compare with two cases: 1) the simulated values and 2) the approximation in [18]. We consider the same transmit power for the source and the relays, i.e., $P = Q$. In Fig. 3.4, we plot the average received SNR values for different transmit powers, where the number of relays is set as 10, 30 and 60. We can see that our approximation is tight for all simulated power values. The approximation in [18], although in closed-form, is largely different from the actual values. Although EVT is exact for large N only, this figure shows that the derived approximation is tight for N as small as 10. In Fig. 3.5, we plot the average received SNR values for networks with different numbers of relays, where the transmit power is set as $P = Q = 10$dB and $P = Q = 15$dB. We can see that the derived closed-form formula is tight for all values of N and the difference with simulated values diminishes as N increases.
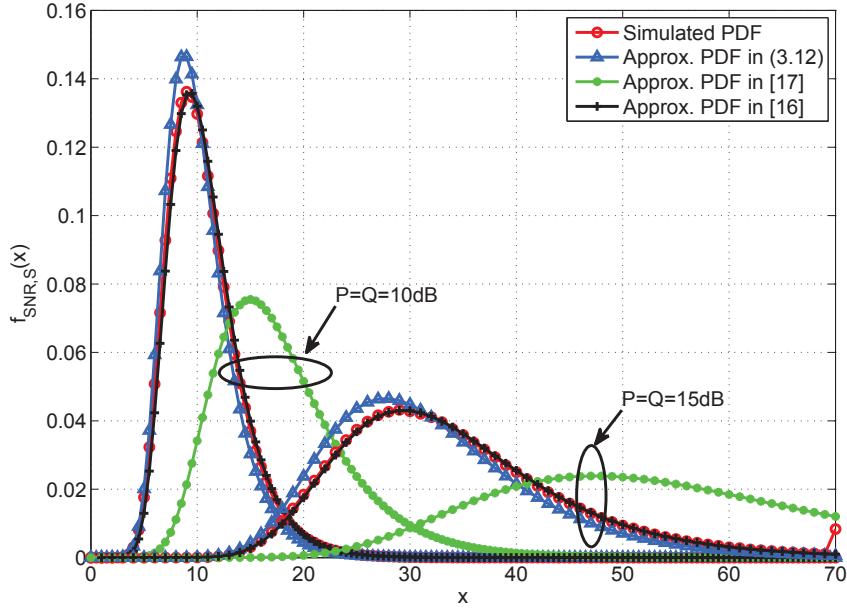
Figure 3.3: PDF of the received SNR for networks with 20 relays.

The approximation in [18] is loose and its error increases as N increase. The curves in this figure also show that the average received SNR is linear in $\log N$ for large N.

In Fig. 3.6, we show the derived ergodic capacity in (3.20) for different power values and compare with the simulated values. The number of relays is set to be 10, 30, and 60. The simulation shows that our approximation is tight for a wide range of N and power values. In the derivation of (3.20), Jensen's inequality is used. So to understand its effect, we also show the ergodic capacity by direct integration using the approximate PDF in (3.12). The values are obtained numerically. We can see that the effect of Jensen's inequality is negligible. We also compare with the minimum link-SNR approximation in (3.21). For the readability of the figure, we only show this result when $N = 30$. We can see that (3.21) is largely deviated from the simulated result.
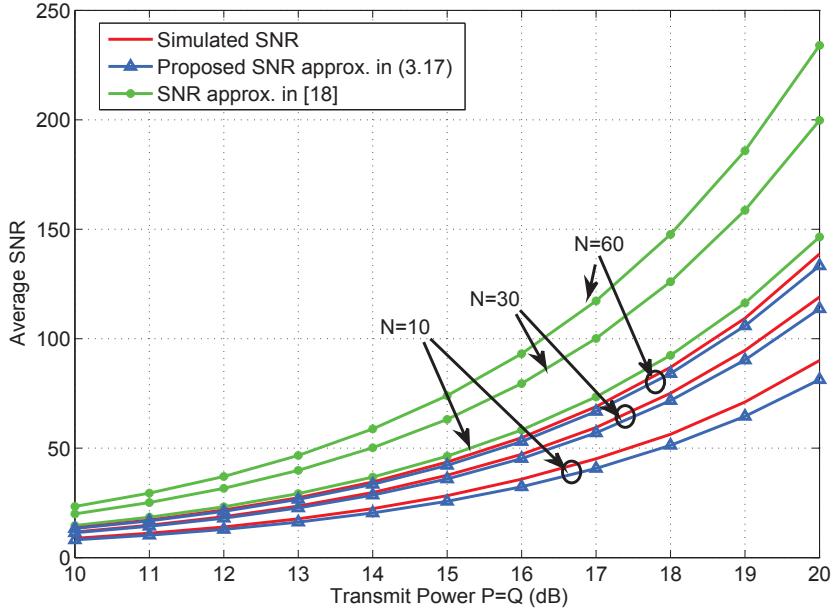
Figure 3.4: Average received SNR for different power in networks with 10, 30, and 60 relays.

3.6 Conclusion

In this chapter, with the help of EVT and manipulations of special functions, closed-form approximations for the average SNR and ergodic capacity of BRS in a large-scale relay network is derived for the high power range. Compared with existing results in integral forms or with special functions, our results are in closed-form and provide useful insights on the behaviour of the array gain and the ergodic capacity with respect to network parameters. Simulation results show that the proposed approximations are tight for a wide range of the transmit power and number of relays, and are significantly superior to existing closed-form results.

$\sim$
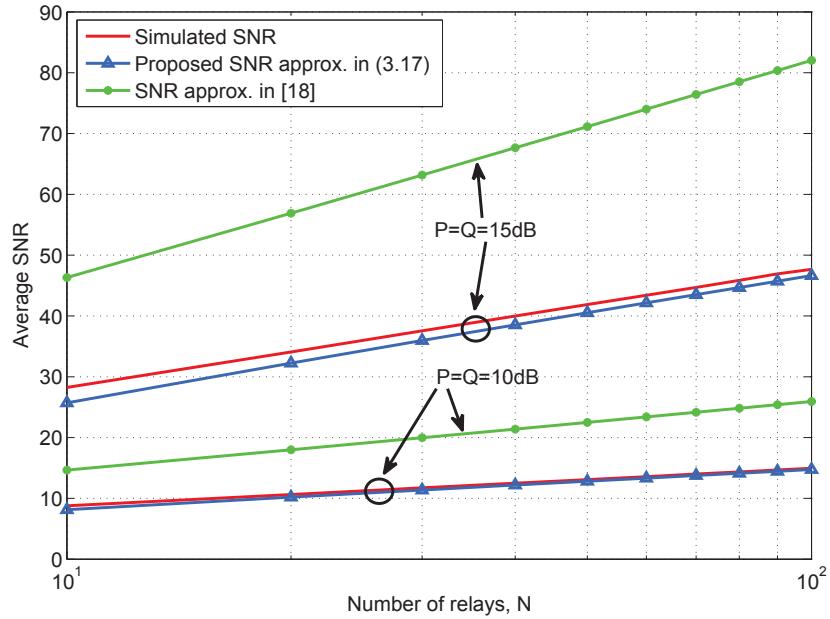
Figure 3.5: Average received SNR for different number of relays in networks with transmit power 10 dB and 15 dB.
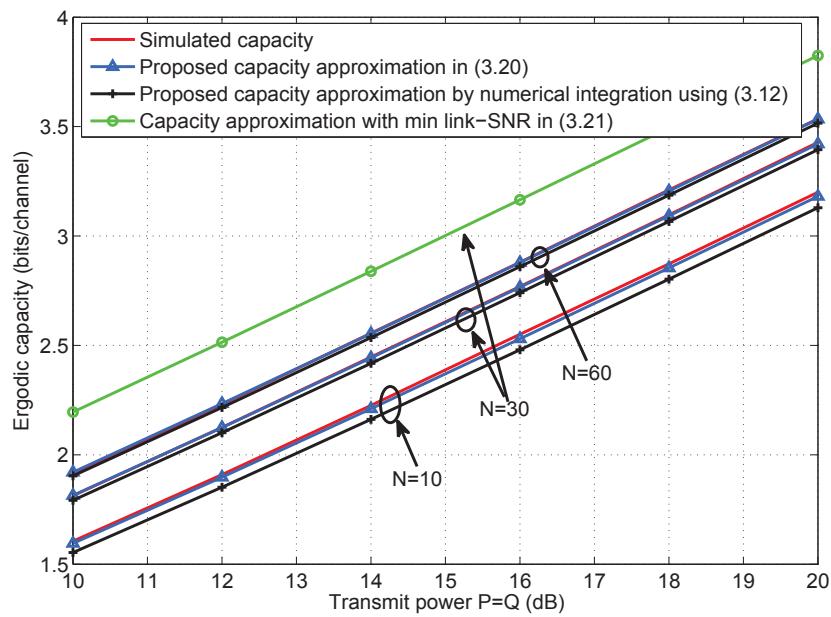
Figure 3.6: Ergodic capacity for different power in networks with 10, 30, and 60 relays.

Chapter 4

Power Allocation and Performance Analysis of Distributed Relay Beamforming in Multiple-User Networks

In this chapter, for a multi-user multi-relay network with a total relay power constraint, we investigate the relay PA that maximizes the network sum-rate. With the optimal relay beamforming for each user, the problem reduces to finding the optimal power the relays use in total to help each user. We first prove that the problem is convex and propose a closed-form suboptimal solution. Further, the SNR and sum-rate are analyzed for networks with a large number of relay antennas. The asymptotic behaviour of the SNR is derived rigorously for the high transmit power regime. Simulation results are provided to show the significance of proper PA and to justify the analytical SNR and sum-rate results.

4.1   Introduction

As introduced in Section 2.3.1, there have been numerous publications on distributed relay beamforming and the PA problem, e.g., [12, 33–35, 47–58, 77–80].

Early research on distributed relay beamforming and relay PA focused on single-user networks, e.g., [12, 33–35, 47, 48, 77–79]. In particular, for single-antenna multi-relay networks with perfect CSI, [35] and [34] provided closed-form solutions for the relay beamforming vector under a total relay power constraint and per-

relay power constraint, respectively. For a multi-antenna single-relay network, the optimal relay beamforming was solved in [77]. [12, 47, 48] investigated distributed beamforming with partial CSI.

To fulfil current and future communication demands, concurrent multi-user transmissions are needed. Multi-user communications can also lead to high bandwidth efficiency. Recently, some work on beamforming and PA in multi-user relay networks was presented [49–58, 80]. In [49], distributed relay PA among sub-channels in a multi-user multi-channel cellular relay network was investigated using non-cooperative game. [50, 51] also used game theory framework to solve the relay PA among users in a multi-user single-relay network. [52, 53] worked on the relay PA in multi-user multi-relay single-antenna networks, where user and relay transmissions are on orthogonal channels. In [57, 58, 80], the optimal relay beamforming designs and PA were investigated for multi-user multi-relay networks with the existence of user-interference. [57, 58] considered single-antenna relays; while [80] was on multi-antenna relays. [54] researched on the optimal relay beamforming design for multi-user multi-relay networks with quantized partial CSI. In [55, 56], beamforming designs for multi-user multi-relay networks, where user transmissions are orthogonal but relay transmissions are non-orthogonal, were considered. In [55], all users transmit to the same destination; while in [56], users transmit to different destinations. In [56], numerical algorithms for relay PA were provided to solve two problems: the total relay power minimization under received SNR constraints and the maximization of the minimum SNR with total and separate relay power constraints.

In this chapter, we consider the same multi-user relay network as in [56], but investigate the relay beamforming and PA that maximizes the network sum-rate. Thus, our problem is different to [56]. Compared with aforementioned references, our work is different in the network model and the transmission/reception techniques. By using the results of optimal beamforming for single-user networks, we first simplify the problem formulation and show that the PA problem is convex. A closed-form approximate solution is also proposed. Further, for a network with asymptotically large number of relay antennas, we analyze the received SNR of

43

each user and the network sum-rate. For high transmit power, we rigorously derive the dominant term of the received SNR formula. To the best of our knowledge, our work is the first that provides theoretical analysis on the performance of distributed relay beamforming; while previous work relied solely on numerical simulations. We should clarify that our analytical results are asymptotic and valid for large number of relay antennas only (applicable to massive-antenna relay stations [10, 36]).

The remaining of the chapter is organized as follows. In the next section, the multi-user relay network model and the relay PA problem are introduced. In Section 4.3, we show that the sum-rate maximizing relay PA problem is convex, and propose a closed-form suboptimal solution. Section 4.4 is on the received SNR and network sum-rate analysis when the number of relay antennas is large. Section 4.5 shows the simulation results and Section 4.6 contains the conclusions.

4.2 Problem Formulation

Consider a one-way multi-user multi-relay network in Fig. 4.1. There are M users, each sending information to its own destination with the help of N relays. Every node is equipped with a single antenna. Denote the channel from User m to Relay n as f_{mn} and from Relay n to Destination m as g_{nm}. There is no direct link between the users and the destinations. The channels are assumed to be i.i.d. and follow the zero-mean CSCG distribution, i.e., Rayleigh fading channels. Assume that f_{mn} has variance $\sigma_{f,m}^2$, and g_{nm} has variance $\sigma_{g,m}^2$. With this, we actually assume that channels corresponding to the same relay but different users/destinations have different variances; while channels corresponding to different relays but the same user/destination have the same variance. This applies to networks where the relay antennas are co-located or the relays are close to each other; while the users are arbitrarily distributed. The transmit power of each user is assumed to be the same, denoted as P. The total relay power is assumed to be Q.

To avoid interference, the transmissions of different source-destination pairs are allocated with orthogonal channels, e.g., frequency division multiple access (FDMA). But the relay transmissions in helping the same user are non-orthogonal.
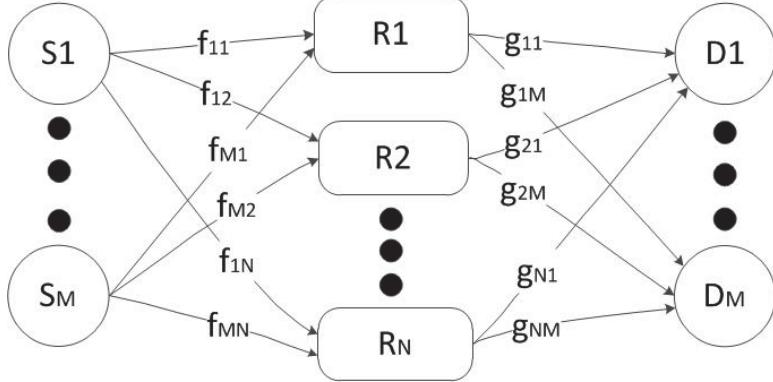
Figure 4.1: Multi-user multi-relay network model.

Every relay helps all users. The transmission and reception of User m's information in the mth channel are as follows. The two-step distributed beamforming scheme [34, 35] is adopted. In Step 1, User m sends its signal s_m to the relays; and in Step 2, all relays adjust the power and phases of their received signals, and forward to Destination m. For the optimal phase adjustment, Relay n should cancel the phases of f_{mn} and g_{nm}, so the forwarded signals add coherently at the destination. For the power adjustment, we use coefficient α_{nm}. That is, the amount of power Relay n uses for User m is $\alpha_{nm}^2 Q$. By following the description in [34, 35], the transceiver equation of User m can be written as

$$x_m = \sum_{n=1}^{N} \frac{\alpha_{nm}|f_{mn}g_{nm}|\sqrt{PQ}}{\sqrt{1+|f_{mn}|^2 P}} s_m + \sum_{n=1}^{N} \frac{\alpha_{nm}|g_{nm}|\sqrt{Q}}{\sqrt{1+|f_{mn}|^2 P}} v_{mn} + w_m, \quad (4.1)$$

where x_m is the received signal at Destination m, v_{mn}, w_m are the noises at the nth relay (with a phase shift) and the destination. They are assumed to be i.i.d. following $\mathcal{CN}(0, 1)$.

From (4.1), the received SNR of User m can be calculated to be

$$\text{SNR}_m = \frac{\sum_{n=1}^{N} \frac{\alpha_{nm}^2 |f_{mn}g_{nm}|^2 PQ}{1+|f_{mn}|^2 P}}{\sum_{n=1}^{N} \frac{\alpha_{nm}^2 |g_{nm}|^2 Q}{1+|f_{mn}|^2 P} + 1}. \quad (4.2)$$

The network sum-rate is thus

$$R_{\text{sum}} = \sum_{m=1}^{M} \log_2 \left(1 + \text{SNR}_m\right).$$

Recall that $\alpha_{nm}^2 Q$ is the power Relay n uses in helping User m. Since the total power of all relays is constrained to Q, we have $\sum_{m=1,n=1}^{M,N} \alpha_{nm}^2 \leq 1$. It is easy

45

to show that the optimal performance is obtained when the constraint takes the equality. The PA problem is thus the following:

$$\max \sum_{m=1}^{M} \log_2 \left(1 + \text{SNR}_m\right) \text{ s.t. } \sum_{m=1,n=1}^{M,N} \alpha_{nm}^2 = 1. \tag{4.3}$$

4.3 Power Allocation Solution

Let $\gamma_m \triangleq \sum_{n=1}^{N} \alpha_{nm}^2$, which is the percentage of the total relay power used for User m. From (4.3), we have $\sum_{m=1}^{M} \gamma_m = 1$. The problem in (4.3) can be re-written as

$$\max \sum_{m=1}^{M} \log_2 \left(1 + \max_{\sum_{n=1}^{N} \alpha_{nm}^2 = \gamma_m} \text{SNR}_m\right)$$

$$\text{s.t. } \sum_{m=1}^{M} \gamma_m = 1 \text{ and } \gamma_m \geq 0, \text{ for any } m.$$

For a fixed γ_m, the sub-problem

$$\max_{\sum_{n=1}^{N} \alpha_{nm}^2 = \gamma_m} \text{SNR}_m$$

is the same as the relay PA in single-user multi-relay networks with total relay power $\gamma_m Q$. We can use the result in [35] to get the optimal solution:

$$\alpha_{nm}^2 = c_m \frac{|f_{mn} g_{nm}|^2 \left(1 + |f_{mn}|^2 P\right) PQ}{1 + |f_{mn}|^2 P + \gamma_m |g_{nm}|^2 Q}, \tag{4.4}$$

$$c_m = \gamma_m \left[\sum_{n=1}^{N} \frac{|f_{mn} g_{nm}|^2 \left(1 + |f_{mn}|^2 P\right) PQ}{1 + |f_{mn}|^2 P + \gamma_m |g_{nm}|^2 Q}\right]^{-1},$$

with which

$$\text{SNR}_m = \gamma_m \sum_{n=1}^{N} \frac{|f_{mn} g_{nm}|^2 PQ}{1 + |f_{mn}|^2 P + \gamma_m |g_{nm}|^2 Q}. \tag{4.5}$$

The problem thus reduces to

$$\max \sum_{m=1}^{M} \log_2 \left(1 + \gamma_m \sum_{n=1}^{N} \frac{|f_{mn} g_{nm}|^2 PQ}{1 + |f_{mn}|^2 P + \gamma_m |g_{nm}|^2 Q}\right)$$

$$\text{s.t. } \sum_{m=1}^{M} \gamma_m = 1 \text{ and } \gamma_m \geq 0, \text{ for any } m. \tag{4.6}$$

The original (MN)-dimensional optimization problem in (4.3) is reduced to an M-dimensional one, which is on the optimization of the power the relays use in total in forwarding each user's information.

Proposition 2. *The optimization problem (4.6) is convex.*

Proof. First, the constraints of the optimization problem are linear, thus convex. By directly calculating the Hessian of the objective function and seeing that it is negative semi-definite, we can show that the objective function is concave. Thus the optimization problem (4.6) is convex. $\qquad\square$

This proposition helps to find the globally optimal PA via efficient convex optimization algorithms such as the interior method [81].

Next we propose an approximate solution in closed-form. Ignoring the γ_m in the denominator, we approximate the SNR in (4.5) as

$$\mathrm{SNR}_m \approx \gamma_m \sum_{n=1}^{N} \frac{|f_{mn}g_{nm}|^2 PQ}{1 + |f_{mn}|^2 P + |g_{nm}|^2 Q}.$$

This approximation is valid when $\gamma_m Q|g_{nm}|^2$ is small compared with the other two terms in the denominator. Thus, it is expected to be tight when $Q \ll 1$ or $P \gg Q$, corresponding to networks where the relays have limited power or significantly less power than the users. These are also scenarios where proper PA among users is crucial. We define

$$h_m \triangleq \sum_{n=1}^{N} \frac{|f_{mn}g_{nm}|^2 PQ}{1 + |f_{mn}|^2 P + |g_{nm}|^2 Q} \tag{4.7}$$

to simplify the presentation. The PA problem is thus approximated as

$$\max \sum_{m=1}^{M} \log_2 \left(1 + \gamma_m h_m\right)$$

$$\text{s.t.} \sum_{m=1}^{M} \gamma_m = 1 \text{ and } -\gamma_m \le 0, \text{ \textit{for any} } m.$$

A closed-form solution[1] can be found by water-filling algorithm within M steps. The typical water-filling algorithm provides the solution as

$$\gamma_i = \max \left(\frac{1}{i} \left(1 + \sum_{j=1}^{i} h_j^{-1}\right) - \frac{1}{h_i}, 0 \right) \forall i, \text{ s.t.} \sum_{m=1}^{M} \gamma_m = 1.$$

To implement it, an algorithm is presented on the next page. Without loss of generality, we assume that h_m's are ordered as $h_1 > h_2 \cdots > h_M$.

---

[1] According to wikipedia, an expression is in closed-form if it can be evaluated in finite number of operations.

Algorithm 1: Approximate PA solution.

4.4 Performance Analysis

The following theorem on the received SNR is proved.

Theorem 1. *Assume that* $N \to \infty$. *Let* $\eta_m \triangleq \frac{Q\sigma_{g,m}^2}{P\sigma_{f,m}^2} \gamma_m$. *When* $P \gg 1$ *but* η_m *is bounded (e.g., Q has the same order as P), the received SNR of User m has the following behaviour:*

$$\frac{1}{N}\mathrm{SNR}_m = P \left[\sigma_{f,m}^2 \frac{\eta_m^3 - 2\eta_m^2 \log_2(\eta_m) - \eta_m}{(\eta_m - 1)^3} + \mathcal{O}\left(\frac{1}{P}\right) \right]. \tag{4.8}$$

Proof. Proof is available in Appendix C. $\qquad\square$

This theorem shows the asymptotic behaviour of the received SNR for the multi-user large-scale relay network. First, we can see from (4.8) that the received SNR of each user increases linearly with N, the number of relay antennas. On the other hand, if the number of relay antennas is increased by a factor of a, to achieve the same end-to-end SNR, we can decrease both the user transmit power and the total relay power by a factor of $1/a$. The network is thus more energy efficient with more relay antennas. Second, we can see that with a fixed N, if P/Q is bounded (powers of the relays and the users have the same scaling), the SNR increases linearly in P, the transmit power. With further manipulation, we can also show that if both P and Q increase but at different rates, the SNR increases linearly in $\min\{P, Q\}$.

From (4.8), when $P, Q \gg 1$, we have

$$\mathrm{SNR}_m \approx NP\sigma_{f,m}^2 \frac{\eta_m^3 - 2\eta_m^2 \log_2(\eta_m) - \eta_m}{(\eta_m - 1)^3} \triangleq \widetilde{\mathrm{SNR}}_m. \tag{4.9}$$

We can also calculate the sum-rate of the network as follows:

$$R_{\text{sum}} = \sum_{m=1}^{M} \log_2\left[NP\sigma_{f,m}^2 \frac{\eta_m^3 - 2\eta_m^2 \log_2(\eta_m) - \eta_m}{(\eta_m - 1)^3} \right] + \mathcal{O}\left(\frac{1}{P}\right)$$

$$\approx \sum_{m=1}^{M} \log_2(1 + \widetilde{\text{SNR}}_m). \tag{4.10}$$

Relay PA that maximizes (4.10) can be conducted. It can be shown that the sum-rate optimization problem is still convex. But compared with (4.6), the sum-rate in (4.10) is independent of the instantaneous channel coefficients, and only depends on the channel variances. So the PA optimization can be conducted off-line and its complexity is no longer a concern, due to the massive number of antennas at the relays.

4.5 Simulation Results
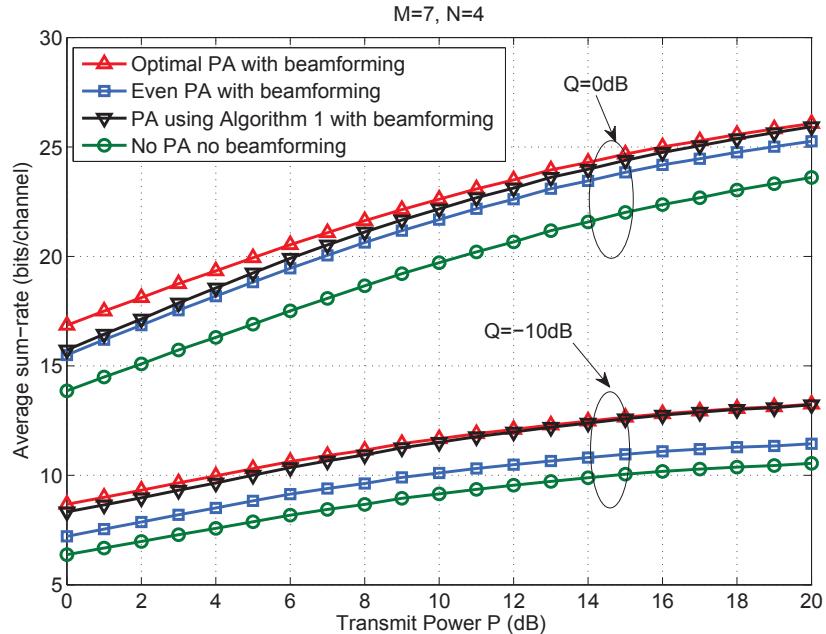
Figure 4.2: Network sum-rate under different PAs.

Figure 4.2 shows the simulated sum-rate of a network with 7 users and 4 relays ($M = 7, N = 4$) under different power conditions. We use a simple network configuration, where all nodes are on one line. The relays are located at the origin;

while the locations of all sources and destinations are generated randomly between $(-1, 1)$ following uniform distribution. A pass-loss exponent 4 is assumed, so the variance of a channel is d^{-4} if the distance between its two nodes is d. The channel magnitudes follow independent Rayleigh distributions. We compare four PA schemes: 1) no PA, where $\alpha_{mn} = 1/(MN)$; 2) even PA, where $\gamma_m = 1/M$ with optimal relay beamforming in (4.4); 3) optimal PA with optimal relay beamforming obtained via convex optimization; and 4) approximate PA with optimal relay beamforming using Algorithm 1. We can see from the figure that proper PA among users and relay beamforming can significantly improve the network sum-rate. The proposed approximate solution (in closed-form) has close performance to the optimal solution when Q is small or $P \gg Q$.
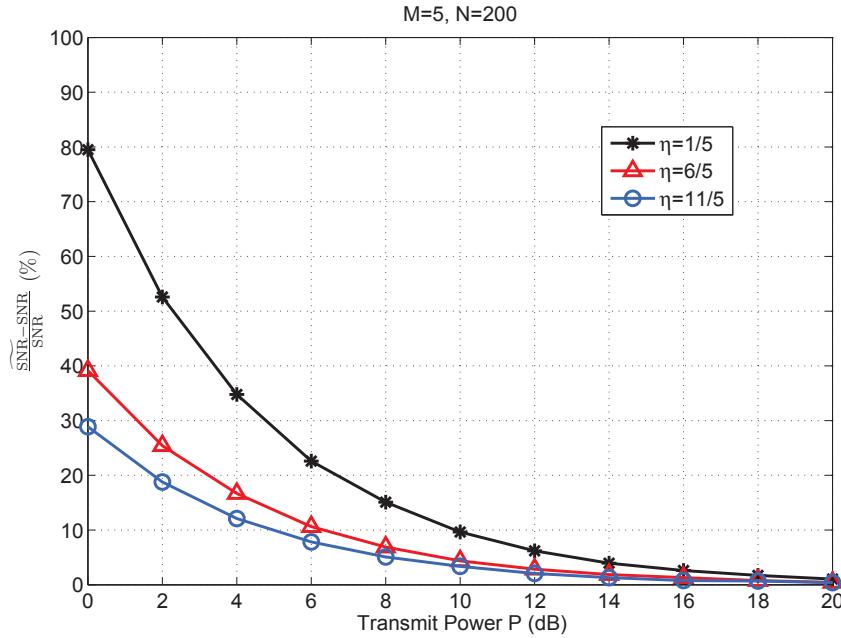
Figure 4.3: Ratio of SNR-difference to SNR.

In Figure 4.3, we consider a network with 5 users and 200 relays ($M = 5, N = 200$). All channel variances are assumed to be the same. The users thus have the same received SNR and the optimal PA is $\gamma_m = 1/M$. Also, $\eta_m = \eta = Q/(PM)$. We show the ratio of the difference of the derived SNR in (4.9) and the true SNR to the true SNR value for different η's, i.e., $\frac{\widetilde{\text{SNR}}-\text{SNR}}{\text{SNR}}$. It can be seen from the figure that the derived SNR is very close to the true SNR value when the transmit power

50

is high.

4.6   Conclusion

In this chapter, for a multi-user multi-relay network, we investigated the PA problem that maximizes the sum-rate. The problem was proved to be convex. An approximate closed-form solution was also proposed. For networks with a large number of relay antennas, we analytically derived the asymptotic behaviour of the user SNR and the network sum-rate. Simulation results were shown to justify the importance of proper PA among users and also to verify the performance analysis results.

$\sim$

Chapter 5

Performance Analysis and Scaling Law of MRC/MRT Relaying with CSI Error in Centralized Relay Networks

This chapter provides a comprehensive scaling law and performance analysis for multi-user massive MIMO relay networks, where the relay station is equipped with a large number of antennas and uses the combination of MRC and MRT for low-complexity processing. CSI error is considered in our model and analysis. First, a sum-rate lower bound is derived which manifests the effect of system parameters including the number of relay antennas, the number of users, the CSI quality, and the transmit powers of the sources and the relay. Via a general scaling model on the system parameters with respect to the number of relay antennas, the asymptotic scaling law of the network sum-rate as a function of the parameter scalings is obtained. Th results show quantitatively the trade-off between the network parameters and their effect on the sum-rate. In addition, a sufficient condition on the parameter scalings for the SINR to be asymptotically deterministic is given, which cover existing studies on asymptotically deterministic analysis as special cases. Then, the scenario where the SINR increases linearly with the number of relay antennas is studied. A sufficient and necessary condition on the parameter scaling for this scenario is proved. It is shown that in this case, the interference power is not asymptotically deterministic. The distribution of the interference power is derived, based on which the outage probability and ABER of the relay network are analysed.

5.1 Introduction

As the performance of massive MIMO networks provides benchmark for the performance of the large-scale relay networks, in this section, the literature on the performance analysis of massive MIMO networks will firstly be briefly reviewed.

With the advantages of massive MIMO (as have been discussed in Section 1.1.2), its performance has been widely studied in the literature [62, 85–89]. In [85], for the uplink of massive MIMO systems with MRC or ZF, the deterministic equivalence of the achievable sum-rate is derived by using the law of large numbers. The CSI is obtained with MMSE estimation. The following power scaling laws are shown. With perfect CSI, the user and/or relay power can be scaled down linearly with the number of antennas while maintaining the same SINR; when there is CSI error and the training power equals the data transmit power, the power can only be scaled down by the square root of the number of antennas. Besides, it is also shown that with MRC or ZF, the spectral efficiency and energy efficiency can be improved simultaneously. Another work on the energy efficiency and power efficiency of a single-cell multi-user massive MIMO network is reported in [62], where a Bayesian approach is used to obtain the capacity lower bounds for both MRT and ZF precodings in the downlink. It is shown that that for high spectral efficiency and low energy efficiency, ZF outperforms MRT, while at low spectral efficiency and high energy efficiency the opposite holds. While the channel models used in [62, 85] are Rayleigh fading, Ricean fading channel is considered in [86] in the uplink of massive MIMO systems, where the CSI is also obtained with MMSE estimator. Sum-rate approximations on the MRC and ZF receivers are obtained using the mean values of the components in the SINR formula. The derived power scaling law is that when the CSI is perfect or Ricean factor is non-zero, the user transmit power can be scaled down inversely proportional with the number of antennas while maintaining the same SINR level. Otherwise, the transmit power can only be scaled down inversely proportional to the square root of the antennas.

While the aforementioned work analyses the sum-rate performance and power scaling law, there are also some work on the distribution and outage probability of

the SINR. In [87], the SINR PDF of MRT precoding is derived in closed-form in the downlink of a single-cell multi-user massive MIMO network. Besides, the asymptotic SINR performance is analysed when the number of users remains constant or scales linearly with the number of antennas. For the same network, in [88], the outage probability of MRT precoding was derived in closed-form. The authors first derive the distribution of the interference power. It's shown that for a reasonable transmit power and limited number of users, the interference power dominates the random property of the SINR. Then the outage probability of the SINR is derived in closed-form. While only small-scale fading is considered in [87, 88], the effects of both small-scale (Rayleigh) fading and large-scale (log-normal) fading are considered in [89]. In this work, the PDF of the SINR of MRC receiver is approximated by log-normal distribution. Then the outage probability is derived in closed-form. The analysis showed that the shadowing effect cannot be eliminated by the use of a large number of antennas.

Intrigued by the massive MIMO, research activities on massive MIMO relay networks are increasing in recent years [37–45, 63, 83, 84]. In [83, 84], for a single-user massive MIMO relay network with co-channel interferences at the relay, the ergodic capacity and outage probability of MRC/MRT and ZF relaying schemes are derived in closed-forms. The more general multiple-user massive MIMO relay networks are analysed in [37–45, 63]. Depending on the structure of the network model, the works can be divided to the following two categories.

In [37–39], a network with multiple single-antenna users, one massive MIMO relay station and one massive MIMO destination is considered. This model applies to the relay-assisted uplink multiple-access network, and the destination is the base station. In [37], it is shown with perfect CSI, and infinite relay and destination antennas, the relay or user transmit power can scale inversely proportional to the number of antennas without affecting the performance. When there is CSI error, the user or relay power can only scale down with the square root of the number of antennas, given that the training power equals the transmit power. The same network is also considered in [38, 39] while the co-channel interference and pilot contamination are considered in [38], and the channel aging effect is considered

in [39]. The effects of these factors on the power scaling are shown therein.

Another type of network is the relay-assisted multi-pair transmission network, where multiple single-antenna sources communicate with their own destinations with the help of a massive MIMO relay [40–45, 63]. In [40, 63], the sum-rates of multi-pair massive MIMO relay network with MRC/MRT and ZF relaying are analyzed for one-way and two-way relaying respectively. Perfect CSI is assumed at the relay station. In both work, with the deterministic equivalence analysis, it is shown that the sum-rate can remain constant when the transmit power of each source and/or relay scales inversely proportional to the number of relay antennas. In [41], the same network model as [40] is considered for MRC/MRT relaying where the number of relay antennas is assumed to be large but finite. The analysis shows that, when the transmit powers of the relay and sources are much larger than the noise power, the achievable rate per source-destination pair is proportional to the logarithm of the number of relay antennas, and is also proportional to the logarithm of the reciprocal of the interferer number. Some typical asymptotic scenarios are also discussed and similar power scaling law results to [40] are obtained. In [42], the full-duplex model is considered for one-way MRC/MRT relaying. A sum-rate lower bound is derived with Jensen's inequality. The effect of loop interference is investigated when the source or relay power scales inversely proportional to the number of relay antennas.

While the above work assume perfect CSI at the relay, recent study has turned to networks with estimation error [43–45], which is more practical and challenging to analyze. In [43, 44], a one-way massive MIMO relay network model is considered, where MMSE estimation is used to obtain the CSI. While [43] uses ZF relaying and assumes that the CSI error exists in both hops, [44] uses MRC/MRT relaying and assumes that the CSI error only exists in the relay-destination hop. In both work, the power scalings of the sources and relay for non-vanishing SINR are discussed. It is assumed in the analysis that the training power equals the data transmission power. Compared with previous power scaling law results, the analysis in [43, 44] are more comprehensive by allowing the power scaling to be anywhere between constant and linearly increasing with the number of relay antennas. In [45], a two-

way MRC/MRT relaying network with CSI error is considered. With deterministic equivalence analysis, it is shown that when the source or relay power scales inversely proportional to the number of relay antennas, the effects of small-scale fading, self-interference, and noise caused by CSI error all diminish.

In this chapter, the performance of MRC/MRT relaying in a one-way massive MIMO relay network with CSI error is investigated. Firstly, by deriving a lower bound on the sum-rate, we investigate the performance scaling law with respect to the relay antenna number for a general setting on the scaling of the network parameters, including the number of source-destination pairs, the CSI quality parameter, and the transmit powers of the sources and the relay. The derived scaling law results show explicitly the trade-off between network parameters and the effect of the parameters to the sum-rate. Then, since deterministic equivalence is an important framework for performance analysis of massive MIMO systems, we derive a sufficient condition on the parameter scales for the SINR to be asymptotically deterministic. The results cover existing literature on deterministic equivalence analysis as special cases. In addition, we consider the scenario where the SINR increases linearly with respect to the relay antenna number for applications with high QoS demands. The sufficient and necessary condition for such scenario is provided. Moreover, it is shown that in this case the interference power does not diminish and dominates the statistical performance of the SINR. By deriving the PDF of the interference power in closed-form, expressions for outage probability and ABER are obtained, and their behaviour with respect to different parameters are discussed.

Our major differences from existing work are summarized as blow.

- Our system model is different from all the mentioned existing work in relaying scheme, CSI assumption, or communication protocol. The work with the closest model is [44], where the CSI error is assumed to exist in the relay-destinations hop only. We use a more general model where CSI error exists in both hops.

- In our sum-rate scaling law analysis, a general model for the scaling of each network parameter, including the number of source-destination pairs, the CSI

56

quality parameter, the transmit powers of the source and the relay, is proposed. In this model, the scale exponent with respect to the relay antenna number can take continuous values from '0' to '1'. In most existing work, only a few discrete values for the power scaling, $0, 1, 1/2$, are allowed. Although [43, 44] also allow continuous exponent values, they constrains the number of sources as constant and the training power equals to the transmit power.

- While in existing work, the asymptotically deterministic equivalence analysis is based on the law of large numbers, we use the quantized measure, squared coefficient of variation (SCV), to examine this property. As the law of large numbers only applies to the summation of i.i.d. random variables, by using the SCV, we can discuss the asymptotically deterministic property of random variables with more complex structures. By examining the SCV of each component of the SINR, we can decide if it is asymptotically deterministic so that it can be approximated with its mean value to simplify the SINR expression.

Based on these features that distinguish our work from existing ones, our unique contributions are listed as below.

1. A general scaling-law that relates the scaling of the user SINR and the scalings of the network parameters are derived. The law provides comprehensive insights and reveals quantitatively the trade-off among different system parameters.

2. We derive a sufficient condition on the parameter scales for the SINR to be asymptotically deterministic. Compared with existing work, where only specific asymptotic cases are discussed, our derived sufficient condition is more comprehensive. It covers the cases discussed in previous works, and also shows more asymptotically deterministic SINR scenarios. Besides, for the SINR to be asymptotically deterministic, the trade-off between different parameter scales is also discussed.

57

3. Through the scaling law results, it is shown that, for practical network scenarios, the average SINR is at the maximum linearly increasing with the number of relay antennas. We prove that the sufficient and necessary condition for it is that all other network parameters remain constant. While the existing work mainly focus on the constant SINR case, this linearly increasing SINR case is not well studied. Our work fills this gap and analyses its outage probability and ABER performance.

The remaining of the chapter is organized as follows. In the next section, the system model including both the channel estimation and data transmission under MRC/MRT relaying are introduced. Then the performance scaling law is analyzed in Section 5.3. In Section 5.4, the asymptotically deterministic SINR case is discussed. The linearly increasing SINR case is investigated in Secion 5.5. Section 5.6 shows the simulation results and Section 5.7 contains the conclusion.

5.2 System Model and Preliminaries for Scaling Law Analysis

We consider a multi-pair relay network with K single-antenna sources $(S_1, \cdots, S_K)$, each transmitting to its own destination. That is, S_i sends information to Destination i, D_i. We assume that the sources are far away from the destinations so that no direct connections exist. To help the communications, a relay station is deployed [8]. The number of antennas at the relay station, M, is assumed to be large, e.g., a few hundreds [37–39, 41–45, 64, 65, 83, 84]. This enables the relay station to apply massive MIMO technologies to achieve high energy and spectral efficiency [42, 43, 45]. In addition, we assume $M \gg K$, because under this condition, the simple linear processing, e.g., MRC/MRT at the relay, can have near optimal performance in massive MIMO systems [90].

Denote the $M \times K$ and $K \times M$ channel matrices of the source-relay and relay-destination links as $\mathbf{F}$ and $\mathbf{G}$, respectively. The channels are assumed to be i.i.d. Rayleigh fading, i.e., entries of $\mathbf{F}$ and $\mathbf{G}$ are mutually independent and every entry follows the CSCG distribution with zero-mean and unit-variance, denoted as

$\mathcal{CN}(0, 1)$. The assumption that the channels are mutually independent is valid when the relay antennas are well separated.

The information of $\mathbf{F}$ and $\mathbf{G}$ is called CSI, which is essential for signal processing in the relaying schemes. In practice, the CSI is estimated through channel training. Due to the existence of noise and interferer, channel estimation cannot be perfect but always contains error. The CSI error is an important issue for massive MIMO systems [43–45, 85, 86]. In what follows, we will first describe our channel estimation model, then the data transmission process and MRC/MRT relaying scheme will be introduced.

5.2.1   Channel Estimation

To combine the received signals from the sources and precode the signals for the destinations, the relay must acquire CSI, the values of $\mathbf{F}$ and $\mathbf{G}$. $\mathbf{F}$ is the uplink channel from the sources to the relay, which can be estimated by letting the sources send pilots to the relay. In small-scale MIMO systems, $\mathbf{G}$ can be estimated by sending pilots from the relay to the destinations and the destinations will feedback the estimation results to the relay [8, 29]. However, this strategy is not viable for massive MIMO systems, as the number of time slots needed for pilots grows linearly with the number of relay antennas M, which may exceed the channel coherence interval. Consequently, to estimate $\mathbf{G}$, we assume a time-division-duplexing (TDD) system with channel reciprocity [10]. So pilots are sent from the destinations and the relay-destination channels can be estimated at the relay station.

Without loss of generality, we elaborate the estimation of $\mathbf{F}$, and the estimation of $\mathbf{G}$ is similar. Since the channel estimation is the same as that in the single-hop MIMO system, we will briefly review it and more details can be found in [8, 29] and references therein. Denote the length of the pilot sequences as τ. For effective estimation, τ is no less than the number of sources K [62,85]. Assume that all nodes use the same transmit power for training, which is denoted as P_t. Therefore, the pilot sequences from all K sources can be represented by a $\tau \times K$ matrix $\sqrt{\tau P_t}\mathbf{\Phi}$, which satisfies $\mathbf{\Phi}^H\mathbf{\Phi} = \mathbf{I}_K$. The $M \times \tau$ received pilot matrix at the relay is

$$\mathbf{Y}_{train} = \sqrt{\tau P_t}\mathbf{F}\mathbf{\Phi}^T + \mathbf{N},$$

59

where N is the $M \times \tau$ noise matrix with i.i.d. $\mathcal{CN}(0, 1)$ elements.

The MMSE channel estimation is considered, which is widely used in the channel estimation of massive MIMO networks [37, 43, 85, 86]. The MMSE estimation of $\mathbf{F}$ given $\mathbf{Y}_{train}$ is

$$\hat{\mathbf{F}} = \frac{1}{\sqrt{\tau P_t}} \mathbf{Y}_{train} \mathbf{\Phi}^* \frac{\tau P_t}{1 + \tau P_t} = \frac{\tau P_t}{1 + \tau P_t} \left(\mathbf{F} + \frac{1}{\sqrt{\tau P_t}} \mathbf{N}_F \right),$$

where $\mathbf{N}_F \triangleq \mathbf{N}\mathbf{\Phi}^*$. As $\mathbf{\Phi}^H \mathbf{\Phi} = \mathbf{I}_K$, $\mathbf{N}_F$ has i.i.d. $\mathcal{CN}(0, 1)$ elements. Similarly, the MMSE estimation of $\mathbf{G}$ is

$$\hat{\mathbf{G}} = \frac{\tau P_t}{1 + \tau P_t} \left(\mathbf{G} + \frac{1}{\sqrt{\tau P_t}} \mathbf{N}_G \right).$$

Define $\mathbf{E}_f \triangleq \hat{\mathbf{F}} - \mathbf{F}$ and $\mathbf{E}_g \triangleq \hat{\mathbf{G}} - \mathbf{G}$ which are the estimation error matrices. Due to the feature of MMSE estimation, $\hat{\mathbf{F}}$ and $\mathbf{E}_f$, $\hat{\mathbf{G}}$ and $\mathbf{E}_g$ are mutual independent. Elements of $\hat{\mathbf{F}}$ and $\hat{\mathbf{G}}$ are distributed as $\mathcal{CN}(0, \frac{\tau P_t}{\tau P_t + 1})$. Elements of $\mathbf{E}_f$ and $\mathbf{E}_g$ are distributed as $\mathcal{CN}(0, \frac{1}{\tau P_t + 1})$. Define

$$E_t \triangleq \tau P_t \text{ and } P_c \triangleq \frac{\tau P_t}{\tau P_t + 1}. \tag{5.1}$$

So E_t is total energy spent in training. P_c is the power of the estimated channel element, representing the quality of the estimated CSI, while $1 - P_c$ is the power of the CSI error. It is straightforward to see that $0 \leq P_c \leq 1$. When $P_c \to 1$, the channel estimation is nearly perfect. When $P_c \to 0$, the quality of the channel estimation is very poor.

Note that, different combinations of τ and P_t can result in the same P_c. For the majority of this chapter, P_c will be used in the performance analysis instead of τ and P_t. This allows us to isolate the detailed training designs and to focus on how the quality of CSI affects the system performance. When we consider special cases with popular training settings, e.g., $\tau = K$ and the same training and data transmission power for users, τ and P_t will be used instead of P_c in modeling the CSI error.

5.2.2 Data Transmissions

With the estimated CSI, the next step is the data transmission. Various relay schemes have been proposed [8]. For massive MIMO systems, the MRC/MRT relaying is

preferred for its computational simplicity, robustness, and high asymptotic performance [40, 41, 44, 45, 63, 83, 84]. In the rest of this section, the data transmission with MRC/MRT relaying and CSI error will be introduced.

Denote the data symbol of S_i as s_i and the vector of symbols from all sources as $\mathbf{s}$. With the normalization $\mathbb{E}(|s_i|^2) = 1$, we have $\mathbb{E}(\mathbf{s}^H \mathbf{s}) = K$, where $(\cdot)^H$ represents the Hermitian of a matrix or vector. Let P be the average transmit power of each source. The received signal vector at the relay is

$$\mathbf{x} = \sqrt{P}\mathbf{F}\mathbf{s} + \mathbf{n_r}, \tag{5.2}$$

where $\mathbf{n_r}$ is the noise vector at the relay with i.i.d. entries each following $\mathcal{CN}(0, 1)$.

With MRC/MRT relaying, the retransmitted signal vector from the relay is $a_e \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \mathbf{x}$, where a_e is the coefficient to normalize the average transmit power of the relay to be Q.

With straightforward calculations, we have

$$a_e^2 = \frac{Q}{\mathbb{E}\{\mathrm{tr}\left((\hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \mathbf{x})(\hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \mathbf{x})^H\right)\}} \approx \frac{Q}{PKP_c^3 M^3(1 + \frac{K}{MP_c} + \frac{1}{PP_c M})}, \tag{5.3}$$

where the approximation is made by ignoring the low order terms of M.

Denote $\mathbf{f}_i$, $\hat{\mathbf{f}}_i$, and $\boldsymbol{\epsilon}_{f,i}$ as the ith columns of $\mathbf{F}$, $\hat{\mathbf{F}}$ and $\mathbf{E}_f$ respectively; $\mathbf{g}_i$, $\hat{\mathbf{g}}_i$ and $\boldsymbol{\epsilon}_{g,i}$ as the ith rows of $\mathbf{G}$, $\hat{\mathbf{G}}$ and $\mathbf{E}_g$ respectively. Then, the received signal at D_i can be written as follows.

$$y_i = a_e \sqrt{P} \mathbf{g}_i \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \mathbf{F}\mathbf{s} + a_e \mathbf{g}_i \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \mathbf{n_r} + n_{d,i},$$

$$= \underbrace{a_e \sqrt{P} \hat{\mathbf{g}}_i \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \hat{\mathbf{f}}_i s_i}_{\text{desired signal}} + \underbrace{a_e \sqrt{P} \sum_{k=1,k\neq i}^{K} \mathbf{g}_i \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \mathbf{f}_k s_k}_{\text{multi-user interference}} + \underbrace{a_e \mathbf{g}_i \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \mathbf{n_r}}_{\text{forwarded relay noise}} +$$

$$\underbrace{a_e \sqrt{P} \boldsymbol{\epsilon}_{g,i} \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \boldsymbol{\epsilon}_{f,i} s_i - a_e \sqrt{P} \hat{\mathbf{g}}_i \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \boldsymbol{\epsilon}_{f,i} s_i - a_e \sqrt{P} \boldsymbol{\epsilon}_{g,i} \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \hat{\mathbf{f}}_i s_i}_{\text{noise due to CSI error}} + n_{d,i}, \tag{5.4}$$

where $n_{d,i}$ is the noise at the ith destination following $\mathcal{CN}(0, 1)$. From (5.4), we can see that the received signal is composed of 5 parts: the desired signal, the multi-user interference, the forwarded relay noise, the CSI error term, and the noise at D_i.

61

Define

$$P_{s,e} \triangleq \frac{|\hat{\mathbf{g}}_i \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \hat{\mathbf{f}}_i|^2}{M^4}, \qquad\qquad P_{i,e} \triangleq \frac{1}{K-1} \sum_{k=1,k\neq i}^{K} \frac{|\mathbf{g}_i \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \mathbf{f}_k|^2}{M^3}, \quad (5.5)$$

$$P_{n,e} \triangleq \frac{\|\mathbf{g}_i \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H\|_F^2}{M^3}, \qquad\qquad P_{e,1} \triangleq \frac{(1-P_c)^2}{M^3} \sum_{n=1}^{K} \sum_{m=1}^{K} \hat{\mathbf{f}}_n^H \hat{\mathbf{f}}_m \hat{\mathbf{g}}_m \hat{\mathbf{g}}_n^H, (5.6)$$

$$P_{e,2} \triangleq (1-P_c) \frac{\|\hat{\mathbf{g}}_i \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H\|_F^2}{M^3}, \quad P_{e,3} \triangleq (1-P_c) \frac{\|\hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \hat{\mathbf{f}}_i\|_F^2}{M^3}. \qquad (5.7)$$

From (5.4) we know that $P_{s,e}, P_{i,e}, P_{n,e}$ and $P_{e,1} + P_{e,2} + P_{e,3}$ are the normalized powers of the signal, the interference, the forwarded relay noise, and the noise due to CSI error respectively. With these definitions, the SINR of the ith source-destination pair can be written as

$$\text{SINR}_i = M \frac{P_{s,e}}{(K-1)P_{i,e} + \frac{1}{P}P_{n,e} + P_{e,1} + P_{e,2} + P_{e,3} + \frac{KP_c^3(1+\frac{K}{MP_c}+\frac{1}{PP_cM})}{Q}}. \qquad (5.8)$$

The achievable rate for the ith source-destination pair is

$$C_i = \mathbb{E}\left\{ \frac{1}{2} \log_2(1 + \text{SINR}_i) \right\}. \qquad (5.9)$$

5.2.3 Preliminaries for Scaling Law Analysis

In the following sections, we study the performance behaviour and asymptotic performance scaling law of the massive MIMO relay network. It is assumed throughout the chapter that the number of relay antennas M is very large and the scaling law is obtained by studying the highest-order term with respect to M.

Due to the complexity of the network, it is impossible to rigorously derive the properties of the SINR and the achievable rate for the general M case. Instead, we find the asymptotic performance properties for very large M with the help of Lindebergy-Lévy CLT. The CLT states that, for two length-M independent column vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, whose elements are i.i.d. zero-mean random variables with variances σ_1^2 and σ_2^2,

$$\frac{1}{\sqrt{M}} \mathbf{v}_1^H \mathbf{v}_2 \xrightarrow{d} \mathcal{CN}(0, \sigma_1^2 \sigma_2^2),$$

where $\xrightarrow{d}$ means convergence in distribution when $M \to \infty$.

Another important concept in the performance analysis of massive MIMO systems is *asymptotically deterministic*. Basically, a random variable sequence X_M with a bounded mean is said to be asymptotically deterministic if it converges in probability to a deterministic value x, i.e.,

$$X_M \xrightarrow{P} x \text{ when } M \to \infty. \tag{5.10}$$

In existing literature, where strong law of large numbers is used to derive the deterministic equivalence, the random variable converges almost surely to a deterministic value. The almost sure convergence implies the convergence in probability [91]. Another type of convergence that implies convergence in probability is the convergence in mean square [91]. For a random variable sequence $X_M, M = 1, 2, \cdots$ with a bounded mean, X_M converges in mean square to a deterministic value x, i.e., $X_M \xrightarrow{m.s.} x$ if

$$\lim_{M \to \infty} \text{Var}\{X_M\} = 0.$$

The convergence in mean square requires the variances of the random variable sequence are asymptotically zero, which is the same as the condition for channel hardening effects [92, 93]. Thus, by the definition of channel hardening effects [92, 93], the convergence in mean square in a massive MIMO system means that the effects of small-scale fading is ignorable when the number of antennas is large. Besides, compared with other types of convergence, the convergence in mean square is more tractable for analysis as only the variances of the variable sequence are needed. Therefore, in this work, we decide a random variable sequence with a bounded mean as asymptotically deterministic when it converges in mean square to a deterministic value.

However, the use of the variance may cause inconvenience and sometimes confusion. One can always scale X_M by $1/M^n$ with large enough n to have the asymptotic deterministic property and the scaled random variable converges in probability to 0. But this does not help the performance analysis when the scaling factor M^n is put back into the SINR formula. Thus to avoid the scaling ambiguity, we use SCV, defined as the square of the ratio of the standard deviation over the mean of the random variable [94].

In this work, we use the variance measure to see whether a random variable is asymptotically deterministic, which is a more direct and less messy method than the definition in (5.10). It is noteworthy that the bounded mean condition is important. Without this condition, the convergence with $M \to \infty$ is not well defined. Thus in this work, a random variable sequence X_M with non-zero mean is said to be asymptotically deterministic if

$$\lim_{M \to \infty} \text{SCV}\{X_M\} = 0. \tag{5.11}$$

This definition is mathematically speaking more strict than (5.10) but is more appropriate for performance analysis in massive MIMO systems.

5.3　Analysis on the Achievable Rate Scaling Law

The general performance scaling law of the massive MIMO relay network will be studied in this section. We start with analysing the components of the received SINR to obtain a large-scale approximation. Consequently, a lower bound on the sum-rate is derived via Jensen's inequality. Then the performance scaling law and conditions for favourable SINR (non-decreasing SINR with respect to M) are derived. Typical network scenarios are discussed. Our analysis will show the relationship between the SINR scale and the parameter scales, and the trade-off between different parameter scales to achieve certain SINR performance.

5.3.1　Sum-Rate Lower Bound and Asymptotically Equivalent SINR

For the analysis of the SINR, we first derive the means and SCVs of the components of the SINR, i.e., $P_{s,e}$, $P_{i,e}$, $P_{n,e}$, $P_{e,1}$, $P_{e,2}$ and $P_{e,3}$.

With the help of CLT and tedious derivations, the following lemma on the means and SCVs of SINR components can be obtained.

Lemma 2.

$$\mathbb{E}\{P_{s,e}\} \approx P_c^4, \qquad\qquad \text{SCV}\{P_{s,e}\} \approx \frac{8}{M}, \qquad (5.12)$$

$$\mathbb{E}\{P_{i,e}\} \approx P_c^3 \left(2 + \frac{K}{MP_c} \right), \qquad (5.13)$$

$$\text{SCV}\{P_{i,e}\} \approx \frac{\frac{4}{K-1} + \frac{8+10P_c}{P_c M} + \frac{K^2+18(K-2)P_c}{(K-1)P_c^2 M^2}}{4 + \frac{K^2}{M^2 P_c^2} + \frac{4K}{MP_c}}, \qquad (5.14)$$

$$\mathbb{E}\{P_{n,e}\} \approx P_c^3 + \frac{K}{M}P_c^2, \qquad \text{SCV}\{P_{n,e}\} \approx \frac{2 + 5P_c - 2P_c^2}{MP_c + \frac{K^2}{MP_c} + 2K}, \ (5.15)$$

$$\mathbb{E}\{P_{e,1}\} \approx \frac{K}{M}P_c^2(1 - P_c)^2, \qquad \text{SCV}\{P_{e,1}\} \approx \frac{3}{K}, \qquad (5.16)$$

$$\mathbb{E}\{P_{e,2}\} = \mathbb{E}\{P_{e,3}\} \approx P_c^3(1 - P_c), \quad \text{SCV}\{P_{e,2}\} = \text{SCV}\{P_{e,3}\} \approx 1. \ (5.17)$$

Proof. The proof is available in Appendix D. $\qquad\qquad\qquad\qquad\qquad\square$

In the lemma, the approximations are made by keeping the dominant terms of M.

With our definitions in (5.5)-(5.7) and by noticing that $P_c \in [0, 1]$, the random variables $P_{s,e}$, $P_{i,e}$, $P_{n,e}$, $P_{e,1}$, $P_{e,2}$, $P_{e,3}$ all have non-zero means. From (5.12), we know that $P_{s,e}$ is asymptotically deterministic since its SCV approaches to 0 as $M \to \infty$. Furthermore, the decreasing rate of its SCV is linear in M, showing a fast convergence rate. Thus, for large M, we can approximate it with its mean values. While for the rest of the components in the SINR, the SCVs depend on the scalings of multiple network parameters (such as K and P_c) and their combinations, which do not necessarily converge to 0. We cannot assume they are asymptotically deterministic so far. With the aforementioned approximation, the SINR expression becomes

$$\text{SINR}_i \approx \frac{MP_c^4}{(K-1)P_{i,e} + \frac{1}{P}P_{n,e} + P_{e,1} + P_{e,2} + P_{e,3} + \frac{KP_c^3(1+\frac{K}{MP_c}+\frac{1}{PP_cM})}{Q}}. \ (5.18)$$

With this simplification, the following result on the sum-rate can be obtained.

Lemma 3. *The achievable rate of User i in the massive MIMO relay network has the following lower bound:*

$$C_i \geq C_{i,LB} \triangleq \frac{1}{2} \log_2 \left(1 + \widetilde{\text{SINR}}_i \right), \qquad (5.19)$$

65

where

$$\widetilde{\text{SINR}}_i \triangleq \frac{1}{\frac{2K}{MP_c} + \frac{K^2}{M^2P_c^2} + \frac{1}{MPP_c} + \frac{K}{M^2PP_c^2} + \frac{K}{MP_cQ} + \frac{K^2}{M^2P_c^2Q} + \frac{K}{M^2PP_c^2Q}}. \quad (5.20)$$

Proof. As $\log_2(1 + \frac{1}{x})$ is a convex function of x [81], according to Jensen's inequality, we have

$$C_i \geq \frac{1}{2}\log_2\left(1 + \frac{1}{\mathbb{E}\left\{\frac{1}{\text{SINR}_i}\right\}}\right).$$

By apply the SINR approximation in (5.18), we have

$$\frac{1}{\mathbb{E}\left\{\frac{1}{\text{SINR}_i}\right\}} = \frac{MP_c^4}{\mathbb{E}\left\{(K-1)P_{i,e} + \frac{1}{P}P_{n,e} + P_{e,1} + P_{e,2} + P_{e,3} + \frac{KP_c^3(1 + \frac{K}{MP_c} + \frac{1}{PP_cM})}{Q}\right\}}$$

$$= \frac{1}{\frac{K-1}{M}\left[\frac{2}{P_c} + \frac{K}{MP_c^2}\right] + \frac{1}{MPP_c} + \frac{K}{M^2PP_c^2} + \frac{K}{M^2}(\frac{1}{P_c} - 1)^2 + \frac{2(1-P_c)}{MP_c} + \frac{K(1 + \frac{K}{MP_c} + \frac{1}{PP_cM})}{MP_cQ}},$$

$$\approx \frac{1}{\frac{2K}{MP_c} + \frac{K^2}{M^2P_c^2} + \frac{1}{MPP_c} + \frac{K}{M^2PP_c^2} + \frac{K}{MP_cQ} + \frac{K^2}{M^2P_c^2Q} + \frac{K}{M^2PP_c^2Q}} = \widetilde{\text{SINR}}_i,$$

where the approximation is made by ignoring the lower order terms of M when $M \gg 1$. Thus the lower bound in (5.19) is obtained. $\square$

Note that, by using the method in Lemma 1 of [86], the sum-rate expression in (5.19) can also be obtained. But with the method in [86], the derived expression is an approximation, while our derivations show that it is a lower bound for large M. On the other hand, from Lemma 1 of [86], we know that the lower bound becomes tighter when the number of relay antennas M or the number of sources K increases. From (5.19) and (5.20), we can see that as expected the achievable rate lower bound increases logarithmically with M and P_c. But its increasing rates with P, Q, $1/K$ are slower than logarithmic increase.

The parameter $\widetilde{\text{SINR}}_i$ has the physical meaning of asymptotic effective SINR corresponding to the achievable rate lower bound. Due to the monotonic relationship in (5.19), to understand the scaling law of the achievable rate is equivalent to understanding the scaling law of $\widetilde{\text{SINR}}_i$. In the next subsection, the scaling law of $\widetilde{\text{SINR}}_i$ with respect to different parameters will be studied.

5.3.2 Scaling-Law Results

Now, the scaling law of the asymptotic effective SINR will be analyzed to show how the system performance is affected by the size of the relay antenna array and other network parameters. To have a comprehensive coverage of network setups and applications, for all system parameters including the number of source-destination pairs K, the source transmit power P, the relay transmit power Q, and the CSI quality parameter P_c, a general scaling model with respect to M is used.

Assume that

$$K = \mathcal{O}(M^{r_k}), \quad \frac{1}{P} = \mathcal{O}(M^{r_p}), \quad \frac{1}{Q} = \mathcal{O}(M^{r_q}), \quad \frac{1}{P_c} = \mathcal{O}(M^{r_c}). \quad (5.21)$$

Thus the exponents r_k, r_p, r_q, and r_c represents the relative scales of K, $1/P$, $1/Q$, and $1/P_c$ with respect to M. For practical ranges of the system parameters, we assume that $0 \leq r_k, r_p, r_q, r_c \leq 1$. The reasons are given in the following.

- The scale of K. Following typical applications of massive MIMO, the number of users should increase or keep constant with the number of relay antennas. Thus $r_k \geq 0$. On the other hand, the increasing rate of K cannot exceed the increasing rate of the number of relay antennas since the maximum multiplexing gain provided by the relay antennas is M. Thus, $r_k \leq 1$.

- The scale of P and Q. Following the high energy efficiency and low power consumption requirements of massive MIMO applications, the source transmit power and relay transmit power should not increase with the number of relay antennas. But they can decrease as the number of relay antennas increases with the condition that their decreasing rates do not exceed the increasing rate of the antenna number. This is because that the maximum array gain achievable from M antennas is M. A higher-than-linear decrease will for sure make the receive SINR a decreasing function of M, which contradicts the promise of massive MIMO communications. This discussion means that $0 \leq r_p, r_q \leq 1$.

- The scale of P_c. From the definition of P_c in (5.1), we have $1/P_c = 1 + 1/E_t$, thus $r_c \geq 0$. This is consistent with the understanding that the CSI quality will

not improve as the number of relay antennas increases, as the training process cannot get benefits from extra antennas [10]. On the other hand, since similar to the data transmission, the total training energy should not has lower scaling than $1/M$, we conclude that $1/P_c$ should not have a higher scaling than M. Thus $r_c \leq 1$.

In the previous modelling of the parameters, the exponents can take any value in the continuous range $[0, 1]$. This is different from most existing work where only one or two special values are assumed for the parameters. Widely used values are $0, 0.5$, and 1, which mean that the parameters scale as a constant, a linear function, and the square-root of M. Our model covers existing work as special cases.

To represent the scaling law of the $\widetilde{\mathrm{SINR}}_i$, define its scaling with respect to M as

$$\widetilde{\mathrm{SINR}}_i = \mathcal{O}(M^{r_s}), \text{ or equivalently, } r_s = \lim_{M \to \infty} \frac{\log_2 \widetilde{\mathrm{SINR}}_i}{\log_2 M}. \tag{5.22}$$

The exponent r_s shows the scaling of $\widetilde{\mathrm{SINR}}_i$.

Theorem 2. *For the massive MIMO relay network with MRC/MRT relaying and CSI error, with the model in (5.21) and (5.22), we have the following performance scaling law:*

$$r_s = 1 - r_c - \max(r_p, r_k + r_q). \tag{5.23}$$

Proof. From (5.20) we can see that, the maximal scaling exponent of the terms in the denominator determines the scaling exponent of $\widetilde{\mathrm{SINR}}_i$ with respect to M. After some tedious calculation, we find that the term with the highest scaling exponent is either $\frac{1}{MPP_c}$ or $\frac{K}{MP_cQ}$. By using the parameter models in (5.21), the results in (5.23) is obtained. $\square$

Sensible massive MIMO system should have $r_s \geq 0$, i.e., asymptotic effective SINR and the sum-rate scale at least as $\mathcal{O}(1)$. Otherwise, the system performance will decrease as M increases, which contradicts the motivations of massive MIMO systems. To help the presentation, we refer to the case where $r_s \geq 0$ as the *favourable-SINR scenario*. The condition for favourable-SINR is presented in the following corollary.

Corollary 1. *The necessary and sufficient condition for the massive MIMO relay network with MRC/MRT relaying and CSI error to have favourable-SINR is*

$$r_c + \max(r_p, r_k + r_q) \leq 1, \quad r_c, r_p, r_q, r_k \in [0, 1]. \tag{5.24}$$

Proof. This is a straightforward extension from (5.23) of Theorem 2. □

The scaling law in (5.23) illustrates quantitatively the concatenation of the scalings of different parameters and their effects on the network performance. The condition in (5.24) forms a region of r_k, r_p, r_q, r_c that makes the SINR favourable. They provide guidelines for the plan and design of the massive MIMO relay network. Next, we discuss the physical meanings of (5.23) and (5.24), and several popular network setups.

Firstly, in (5.23), r_k and r_q appears as a summation. According to their definitions in (5.21), the summation is the scaling exponent of K/Q. Then in (5.23), $\max(r_p, r_k + r_q)$, which also equals $\min(-r_p, -r_k - r_q)$, is the minimum of the power scaling exponents of P and Q/K. Recall that P is the per-source transmit power. And Q/K is the average relay power allocated to each source-destination pair. Thus, from (5.23), we can see that the performance scaling of the SINR is determined by two factors: 1) $\max(r_p, r_k + r_q)$, which is the worse per-source-destination-pair power scaling of the two steps, and 2) P_c, which is the CSI quality.

Further, (5.23) shows that r_s, which represents the scale of the system performance is a decreasing function of both $\max(r_p, r_k + r_q)$ and r_c. Thus high transmit power and better CSI quality result in improved performance. There is a natural tradeoff between the worse per-source-destination-pair power and channel training (e.g., between the data transmission phase and the training phase), and one can compensate for the other to keep the performance scaling of the system unchanged. For the two-step communication, it is the worse step that dominates the overall system performance.

The condition in (5.24) implies $r_k + r_q \leq 1$, which means that for the SINR to be favourable, the scaling of the per-source-destination-pair relay power should be no less that $1/M$. This also shows a trade-off between r_k and r_q. Recall that

$0 \leq r_k, r_q \leq 1$. That is, with extra relay antennas, we can serve more users or use less relay power for the same level of performance, but the improvement in the two aspects has a total limit. For example, two cases satisfying the constraint are 1) $r_k = 1$, $r_q = 0$; 2) $r_q = 1$, $r_k = 0$. The first case means that, when the number of users increases linearly with the number of relay antennas (i.e., $r_k = 1$), the relay power must remain constant (i.e., $r_q = 0$), and thus the goal of saving relay power cannot be achieved. The second case is the opposite: when the relay power is scaled inversely proportional to the number of relay antennas, the goal of serving more users cannot be achieved.

5.3.3   Discussions on Several Popular Network Settings

In this subsection, we further elaborate the scaling law in (5.23) and the condition in (5.24) for popular network settings.

1. First, we consider the case of $r_c = 0$, corresponding to perfect or constant CSI quality case (for example, P_t and τ are large constants). From (5.23) and (5.24), the resulting SINR scaling exponent is $r_s = 1 - \max(r_p, r_k + r_q)$ and the necessary and sufficient condition for favourable SINR is $r_k + r_q \leq 1$. Its physical meaning is that, when the CSI is perfect and for the SINR to be favourable, the most power-saving design is to make both the per-source-destination-pair power of the two hops decrease linearly with the number of antennas. Thus, when the CSI quality is good, we can design the networks to serve more users and/or save power consumption, while maintain certain QoS.

2. Next, we consider the case of $r_c = 1$, which is equivalent to $E_t = \mathcal{O}(1/M)$. This means that the total energy used in training is inversely proportional to the relay antenna number. In this case, the SINR scaling exponent is $r_s = -\max(r_p, r_k + r_q)$. To have favourable SINR, from (5.24), we need $r_p = r_k = r_q = 0$. That is, the source data transmit power, the per-source-destination-pair relay power, and the number of users should all remain constant for favourable SINR when the CSI quality is inversely proportional to

70

the relay antenna number. This shows that the CSI quality is key to the performance of massive MIMO relay networks. With low CSI quality, all the promising features of the massive MIMO network are gone.

3. For any general $r_c \in (0,1)$, favourable SINR requires $\max(r_p, r_k + r_q) \leq 1 - r_c$. That is, the worse transmit power per source-destination pair of the two steps cannot be lower than $\mathcal{O}(1/M^{1-r_c})$. This shows the trade-off between the training phase and the data transmission phase. Also shows how CSI quality affects the system performance.

4. For the most power saving setting where $r_p = 1$ or $r_k + r_q = 1$, the per-pair transmit power of either the two steps scales as $1/M$. To have favourable SINR, $r_c = 0$ is needed. Thus, the per source-destination pair data transmit power of either or both steps can be made inverse proportional to the number of relay antennas. But at the same time, the training power must at least remain constant, not a decreasing function of M. If furthermore $r_k = 0$ (the number of source-destination pairs K remains constant), we have for this setting P or Q scales with $1/M$, which is the major power scaling scenario considered in the literature. It is obvious that our result covers this case, and shows more insights by considering the scales of K and P_c.

5. While in the previous discussions, r_c is treated as a free parameter, next, we consider the special case of $P_t = P$ and $\tau = K$. The condition $P_t = P$ corresponds to the practical scenario that user devices always use the same transmit power, no matter for training or data transmission. It is a common assumption in the literature [43–45]. $\tau = K$ is the minimum training length for effective communication [85]. It is shown in [62] that, for maximal-ratio processing, this is the case when the maximal spectral efficiency is achieved. We can see that, in this case, $r_c = \max\{0, r_p - r_k\}$. Consequently, the SINR scale exponent is $r_s = 1 - \max\{0, r_p - r_k\} - \max(r_p, r_k + r_q)$. For the SINR to be favourable, we need $\max(r_p, r_k + r_q) \leq 1$, $2r_p - r_k \leq 1$, and $r_p + r_q \leq 1$. $r_p = r_c \leq 1/2$. Further, we consider a special case of $r_k = 0$, i.e., the number of source-destination pairs is constant. For favourable SINR

71

we need $r_p \leq 1/2$, i.e., the source transmit power can be reduced by $1/\sqrt{M}$ at maximum. This is same as the conclusion as in [43–45]. But note that our model is different from [43–45] and is more general.

6. Another popular setting is to have the number of source-destination pairs increase linearly with M, i.e., $r_k = 1$. One example is assuming that K/M is a constant as M increases. From (5.23) and (5.24), for this case, the SINR scaling exponent is $r_s = -r_c - r_q$ and to have favourable SINR, we need $r_c = r_q = 0$. Thus, to support such number of source-destinations, the CSI quality must be high and at the same time the relay power cannot decrease with M.

5.4   Systems with Asymptotically Deterministic SINR

One important concept appearing in massive MIMO systems is the asymptotically deterministic property. For example, with receiver combining and/or pre-coding at the base station or relay station, random variables such as the signal power and interference power which are random in finite-dimension cases converge in probability to deterministic values as the number of relay antennas is large [64, 65]. This effect is also called channel hardening [10, 92]. With channel hardening, the small-scale fading effect is negligible, and so is the channel variance in the frequency domain. This not only simplifies many design issues but also enables performance analysis via the deterministic equivalences of the random variables. Many performance analysis results in massive MIMO literature rely on asymptotically deterministic property, e.g., [64, 65, 85].

One important question is thus when the massive MIMO system have asymptotically deterministic SINR for the corresponding performance analysis to be valid. In this section, we derive a sufficient condition on asymptotically deterministic SINR and discuss typical scenarios for it, which include scenarios in existing work as special cases.

Firstly, The SINR expression in (5.8) can be reformulated as

$$\text{SINR}_{i,e} = \frac{M^{r_s} P_{s,e}/P_c^4}{P_{i,e}\frac{K-1}{P_c^4 M^{1-r_s}} + \frac{1}{PP_c^4 M^{1-r_s}}P_{n,e} + \frac{1}{P_c^4 M^{1-r_s}}(P_{e,1}+P_{e,2}+P_{e,3}) + \frac{K(1+\frac{K}{MP_c}+\frac{1}{PP_c M})}{QP_c M^{1-r_s}}}. \tag{5.25}$$

The received SINR is asymptotically deterministic when its SCV approaches zero as $M \to \infty$. However, due to the complex structure of the SINR expression, it is highly challenging to obtain its SCV directly. Alternatively, as is shown in Section 5.3, $P_{s,e}/P_c^4$ is asymptotically deterministic, thus for the SINR to be asymptotically deterministic, the sufficient and necessary condition is that the denominator of the formula in (5.25) is asymptotically deterministic. One sufficient condition is that the SCV of the denominator denoted as SCV_d, is no larger than E/M, for some constant E [1]. By the definition of SCV, this condition can be expressed as

$$\text{SCV}_d = \frac{\text{Var}\left\{P_{i,e}\frac{K-1}{P_c^4 M^{1-r_s}} + \frac{1}{PP_c^4 M^{1-r_s}}P_{n,e} + \frac{1}{P_c^4 M^{1-r_s}}(P_{e,1}+P_{e,2}+P_{e,3})\right\}}{\left(\mathbb{E}\left\{P_{i,e}\frac{K-1}{P_c^4 M^{1-r_s}} + \frac{1}{PP_c^4 M^{1-r_s}}P_{n,e} + \frac{1}{P_c^4 M^{1-r_s}}(P_{e,1}+P_{e,2}+P_{e,3})\right\}\right)^2} \leq \frac{E}{M}. \tag{5.26}$$

From (5.25), we have

$$\frac{P_{s,e}/P_c^4}{P_{i,e}\frac{K-1}{P_c^4 M^{1-r_s}} + \frac{1}{PP_c^4 M^{1-r_s}}P_{n,e} + \frac{1}{P_c^4 M^{1-r_s}}(P_{e,1}+P_{e,2}+P_{e,3}) + \frac{K(1+\frac{K}{MP_c}+\frac{1}{PP_c M})}{QP_c M^{1-r_s}}} = \mathcal{O}(1),$$

and since $P_{s,e}/P_c^4 \xrightarrow{m.s.} 1$, we can see that

$$\mathbb{E}\left\{P_{i,e}\frac{K-1}{P_c^4 M^{1-r_s}} + \frac{1}{PP_c^4 M^{1-r_s}}P_{n,e} + \frac{1}{P_c^4 M^{1-r_s}}(P_{e,1}+P_{e,2}+P_{e,3})\right\} = \mathcal{O}(1).$$

Thus (5.26) is equivalent to that

$$\text{Var}\left\{P_{i,e}\frac{K-1}{P_c^4 M^{1-r_s}} + \frac{1}{PP_c^4 M^{1-r_s}}P_{n,e} + \frac{1}{P_c^4 M^{1-r_s}}(P_{e,1}+P_{e,2}+P_{e,3})\right\} \leq \frac{E'}{M}, \tag{5.27}$$

for some constant E'.

Lemma 4. *A sufficient condition for (5.27) is that the variance of each term in (5.27) scales no larger than $1/M$, i.e., the maximum scale order of* $\text{Var}\left\{P_{i,e}\frac{K-1}{P_c^4 M^{1-r_s}}\right\}$,

---

[1]Note that, when $M \to \infty$, given any positive number α, $1/M^\alpha \to 0$. But for practical applications of the deterministic equivalence analysis in large but finite-dimension systems, we consider the scenario that the SCV decrease linearly with the number of antennas or faster. The derived condition is thus sufficient but not necessary.

$\text{Var}\left\{\frac{1}{PP_c^4 M^{1-r_s}}P_{n,e}\right\}$, $\text{Var}\left\{\frac{1}{P_c^4 M^{1-r_s}}P_{e,1}\right\}$, $\text{Var}\left\{\frac{1}{P_c^4 M^{1-r_s}}P_{e,2}\right\}$, *and* $\text{Var}\left\{\frac{1}{P_c^4 M^{1-r_s}}P_{e,3}\right\}$,
is no larger than $1/M$.

Proof. The variance of $P_{i,e}\frac{K-1}{P_c^4 M^{1-r_s}} + \frac{1}{PP_c^4 M^{1-r_s}}P_{n,e} + \frac{1}{P_c^4 M^{1-r_s}}(P_{e,1} + P_{e,2} + P_{e,3})$
is the summation of two parts: the variances of each term, and the covariance of
every two terms. Now, we will prove that if the variances of each term scales no
larger than $1/M$, their covariance also scales no larger than $1/M$.

To make it general and clear, we define $Y = \sum_{n=1}^{N} X_n$, where N is a limited
constant integer and X_n's are random variables. Without loss of generality, we
assume that $\text{Var}\{X_1\}$ has the highest scale among all $\text{Var}\{X_n\}$'s and $\text{Var}\{X_1\} = \mathcal{O}(1/M^\alpha)$, where $\alpha \geq 1$. The variance of Y is

$$\text{Var}\{Y\} = \sum_{n=1}^{N}\text{Var}\{X_n\} + \sum_{i\neq j}\text{Cov}\{X_i, X_j\}.$$

By the definition of covariance, $\sum_{i\neq j}\text{Cov}\{X_i, X_j\}$ takes the maximum value
when X_n's are linearly correlated, i.e., $X_1 = X_2/a_2 = X_3/a_3 \cdots = X_N/a_N$. In
this case, we can obtain that

$$\sum_{i\neq j}\text{Cov}\{X_i, X_j\} = \text{Var}\{X_1\}\sum_{i\neq j}a_i a_j,$$

where we define $a_1 = 1$.

As $\text{Var}\{X_1\}$ has the highest scale, we have a_n scales no higher than $\mathcal{O}(1)$,
that is, there exists constants c_n's such that $a_n \leq c_n$. Thus $\sum_{i\neq j}\text{Cov}\{X_i, X_j\} = \mathcal{O}(1/M^\alpha)$, and consequently $\text{Var}\{Y\}$ scales no higher than $1/M^\alpha$. $\square$

Given Lemma 4, we only need to find the condition for the variances of $(K - 1)P_{i,e}/(P_c^4 M^{1-r_s})$, $P_{n,e}/(PP_c^4 M^{1-r_s})$, $P_{e,1}/(P_c^4 M^{1-r_s})$, $P_{e,2}/(P_c^4 M^{1-r_s})$, and $\frac{1}{P_c^4 M^{1-r_s}}P_{e,3}$
to scale no larger than $1/M$. Using the results on the variances of SINR compo-

nents, the variances of the terms can be obtained as

$$
\mathrm{Var}\{\frac{K-1}{P_c^4 M^{1-r_s}}P_{i,e}\} = \frac{(K-1)^2}{P_c^2 M^{2-2r_s}}\left(\frac{4}{K-1}+\frac{8+10P_c}{P_c M}+\frac{K^2+18(K-2)P_c}{(K-1)P_c^2 M^2}\right)
$$
$$
\sim \mathcal{O}\left(M^{-(2-2r_s-2r_c-r_k)}\right),
$$
$$
\mathrm{Var}\{\frac{1}{PP_c^4 M^{1-r_s}}P_{n,e}\} = \frac{\frac{2}{P_c^3}+\frac{5}{P_c^2}-\frac{2}{P_c}}{M^{3-2r_s}P^2} \sim \mathcal{O}\left(M^{-(3-2r_s-3r_c-2r_p)}\right),
$$
$$
\mathrm{Var}\{\frac{1}{P_c^4 M^{1-r_s}}P_{e,1}\} = \frac{3K}{M^{4-2r_s}}(\frac{1}{P_c}-1)^4 \sim \mathcal{O}\left(M^{-(4-2r_s-4r_c-r_k)}\right),
$$
$$
\mathrm{Var}\{\frac{1}{P_c^4 M^{1-r_s}}P_{e,2}\} = \mathrm{Var}\{\frac{1}{P_c^4 M^{1-r_s}}P_{e,3}\} = \frac{1}{M^{2-2r_s}}(\frac{1}{P_c}-1)^2
$$
$$
\sim \mathcal{O}\left(M^{-(2-2r_s-2r_c)}\right),
$$

where the scaling behaviour at the end of each line is obtained from the definitions of the scaling exponents in (5.21) and considering the constraints in (5.24). Then, we can see that the condition for the order each term to be no higher than $1/M$ is that both following constrains are satisfied.

$$
r_k + 2r_c + 2r_s \le 1, \quad 2r_p + 3r_c + 2r_s \le 2. \tag{5.28}
$$

Combining constrains (5.24) and (5.28), we get a sufficient condition for the SINR to be deterministic in the following proposition.

Proposition 3. *When $M \gg 1$, a sufficient condition for the SINR to be asymptotically deterministic is*

$$
1)\ r_s + r_c + \max\{r_p, r_k + r_q\} = 1 \quad 2)\ 2r_s + 2r_c + r_k \le 1,
$$
$$
3)\ 2r_s + 3r_c + 2r_p \le 2, \quad 4)\ r_c, r_p, r_q, r_k \in [0, 1].
$$

From constraint 2) of the condition, we can see that $r_s \le 1/2$, meaning that the highest possible SINR scaling is $1/\sqrt{M}$ for the sufficient condition. In addition, $r_c \le 1/2$, meaning that to make the SINR asymptotically deterministic, the CSI quality should scale no lower than $1/\sqrt{M}$. By the definition of P_c in (5.1), the lowest scaling the training power P_t can have is $1/\sqrt{M}$. Note that, for a favourable SINR, the scale of the CSI quality just has to be larger than $1/M$. Therefore, for asymptotically deterministic SINR, the constraint on the CSI quality is more strict.

In what follows, we will investigate typical scenarios for the SINR scaling, which include all possible cases if r_s and r_c are allowed to take values from $\{0, 1/2, 1\}$ only. The trade-off between parameters will be revealed.

1. To achieve both $r_s = 1/2$ (the SINR increases linearly with $\sqrt{M}$) and asymptotically deterministic SINR, the sufficient condition reduces to $r_k = 0$, $r_c = 0$, and $\max\{r_p, r_q\} = 1/2$. It means that when the number of users and the CSI quality remain constant, the lower of the source power and the relay power must scale as $1/\sqrt{M}$. While in existing work, only constant SINR case ($r_s = 0$) has been considered [64,65,85], our result shows that the SINR can scale as $\sqrt{M}$ with asymptotically deterministic property.

2. To achieve $r_s = 0$ (constant SINR level) and asymptotically deterministic SINR, two cases may happen: a) $r_c = 0$ and $\max\{r_p, r_k + r_q\} = 1$; and b) $r_c = 1/2$, $r_k = 0$, $r_p \leq 1/4$ and $r_q = 1/2$.

   For Case a), when the CSI quality has constant scaling (e.g., perfect CSI or high quality channel estimation), the scale of the lower per-pair transmission power of the two hops should scale as $1/M$ for asymptotically constant SINR. This is the case considered in [64,65]. Similar scenarios for massive MIMO systems without relays have also been reported in [85].

   Case b) indicates that when the CSI quality scales as $1/\sqrt{M}$ (e,g., the training power scales as $1/\sqrt{M}$ with fixed training length), the number of source-destination pairs should remain constant, the relay power should scale as $1/M$, and the source power can scale smaller than $1/\sqrt[4]{M}$.

5.5 Systems with Linearly Increasing SINR

In our asymptotically deterministic SINR analysis, the scale of the SINR is no larger than $\mathcal{O}(\sqrt{M})$. While, it can be seen from the SINR scaling formula in (5.23) that, the maximum scale of the SINR with respect to the number of relay antennas M, is $\mathcal{O}(M)$, i.e., linearly increasing with M. In this case the sum-rate scales as $\log_2 M$. This is a very attractive scenario for massive MIMO relay networks, in the sense

that, when $M \gg 1$ significant improvement in the network throughput and communication quality can be achieved. Possible applications for such scenario are networks with high reliability and throughput requirement such as industrial wireless networks and high-definition video.

In this section, we study networks with linearly increasing SINR. First, the condition on the parameter scaling for the SINR to be linearly increasing is investigated. Then we show that in this case the interference power of the system is not asymptotically deterministic, but with a non-diminishing SCV as $M \to \infty$. Thus existing deterministic equivalence analysis does not apply and the small-scale effect needs to be considered in analyzing the performance. We first derive a closed-form PDF of the interference power, then obtain expressions for the outage probability and ABER. Their scalings with network parameters are revealed.

Linearly increasing SINR means that the SINR scaling exponent is 1, i.e., $r_s = 1$. Thus the SINR can be formulated as

$$\text{SINR}_{i,e} = M \frac{P_{s,e}/P_c^4}{P_{i,e}\frac{K-1}{P_c^4} + \frac{1}{PP_c^4}P_{n,e} + \frac{1}{P_c^4}(P_{e,1} + P_{e,2} + P_{e,3}) + \frac{K(1+\frac{K}{MP_c}+\frac{1}{PP_cM})}{QP_c}}.$$

From SINR scaling expression in (5.23), we can see that the sufficient and necessary condition for $r_s = 1$ is $r_c = r_p = r_k = r_q = 0$ (note that $r_c, r_p, r_q, r_k \in [0,1]$), i.e., the CSI quality, the source transmit power, the relay power, and the number of users all remain constant.

With the parameter values, we can calculate that the SCVs of $P_{s,e}/P_c^4$ and $P_{n,e}/P/P_c^4$ scales of $1/M$. Therefore, they are asymptotically deterministic and can be approximated with their mean values. On the other hand, the SCVs of $(K-1)P_{i,e}/P_c^4$, $P_{e,1}/P_c^4$, $P_{e,2}/P_c^4$, and $P_{e,3}/P_c^4$ are constant. But we can see that

$$\text{Var}\{\frac{K-1}{P_c^4}P_{i,e}\} \geq \frac{4(K-1)}{P_c^2} \geq 4(K-1)\left(\frac{1}{P_c}-1\right)^2 = 4(K-1)\text{Var}\{\frac{P_{e,2}}{P_c^4}\}.$$

Notice that $P_{e,2}$ and $P_{e,3}$ have the same distribution. As we mainly consider the non-trivial case that $K \geq 3$, we have $\text{Var}\{(K-1)P_{i,e}/P_c^4\} \gg P_{e,2}/P_c^4, P_{e,3}/P_c^4$, especially when the CSI quality P_c is high. Besides, the mean of $P_{e,1}/P_c^4$ scales as $1/M$, and its variance scales as $1/M^2$. Thus the variance of this term is also much smaller than $P_{i,e}(K-1)/P_c^4$. Therefore, $P_{i,e}(K-1)/P_c^4$ dominates the random

behaviour of the SINR and other terms can all be approximated with their mean values. Thus the SINR can be approximated as

$$\text{SINR}_{i,e} \approx \frac{M}{P_{i,e}\frac{(K-1)}{P_c^4} + \left(\frac{1}{P} + \frac{K}{Q}\right)\left(\frac{1}{P_c} + \frac{K}{MP_c^2}\right) + 2\left(\frac{1}{P_c} - 1\right) + \frac{K}{M}\left(\frac{1}{P_c} - 1\right)^2},$$

(5.29)

where only dominant terms of M are kept.

Now, we can conclude that for linearly-increasing SINR, the interference power is not asymptotically deterministic and does not diminish as M increases. In addition, the randomness of the interference power is the dominant contributor to the random behaviour of the SINR. With this result, to analyse the outage probability and ABER performance, the distribution of the interference needs to be derived.

Proposition 4. *Define*

$$\rho_e = \frac{1}{\sqrt{M}}\frac{\sqrt{\frac{4}{P_c} + 10}}{2 + \frac{K}{MP_c}},$$

(5.30)

$$b_e = (K-1)\rho_e, \quad c_e = 1 - \rho_e, \quad d_e = \frac{P_c^3}{K-1}\left(2 + \frac{K}{MP_c}\right).$$

(5.31)

When $M \gg 1$, the PDF of $P_{i,e}$ has the following approximation:

$$f_{P_{i,e}}(y) = \frac{c_e}{b_e + c_e}\sum_{i=0}^{\infty}\left(\frac{b_e}{b_e + c_e}\right)^i \phi(y; K+i-1, d_e c_e),$$

(5.32)

where $\phi(y; \alpha, \beta) = \frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha(\alpha-1)!}$ is the PDF of Gamma distribution with shape parameter α and scale β. It can also be rewritten into the following closed-form expression:

$$f_{P_{i,e}}(y) \approx \frac{(b_e + c_e)^{K-3}}{d_e b_e^{K-2}} \times$$

$$\left[\exp\left(-\frac{y}{d_e(b_e + c_e)}\right) - \exp\left(-\frac{y}{d_e c_e}\right)\sum_{n=0}^{K-3}\frac{1}{n!}\left(\frac{b_e}{d_e c_e(b_e + c_e)}y\right)^n\right].$$

(5.33)

Proof. The proof is available in Appendix E. □

From (5.32), it can be seen that the distribution of the interference power is a mixture of infinite Gamma distributions with the same scale parameter which is $d_e c_e$ but different shape parameters. But as (E.3) is in the form of an infinite summation,

78

it is manipulated to be (5.33) for further analysis. Besides, when the CSI quality is good, i.e., P_c is close to 1, we have $K/(MP_c) \ll 1$ and thus ρ_e and d_e can be simplified by ignoring the term $K/(MP_c)$. Compared with the perfect CSI case where $P_c = 1$, the CSI error makes $d_e c_e$, the scale parameter, smaller.

With this distribution, the outage probability and ABER of the network can be derived as shown in the next two subsections.

5.5.1   Outage Probability Analysis

Outage probability is defined as the probability that the SINR falls below a certain threshold. Due to the complexity of relaying communications, the user-interference and the large scale, the outage probability analysis of multi-user massive MIMO relay networks is not available in the literature. With the help of CLT, we obtained an approximate PDF for the interference power and derived simplified SINR approximation in (5.29) for linearly increasing SINR case. These allow the following outage probability derivation.

Let γ_{th} be the SINR threshold and define

$$\xi \triangleq \left(\frac{1}{P} + \frac{K}{Q} \right) \left(\frac{1}{P_c} + \frac{K}{MP_c^2} \right) + 2 \left(\frac{1}{P_c} - 1 \right) + \frac{K}{M} \left(\frac{1}{P_c} - 1 \right)^2.$$

The outage probability of User i can be approximated as

$$P_{out}(\gamma_{th}) = \mathbb{P}(\mathrm{SINR}_{i,e} < \gamma_{th})$$

$$\approx \mathbb{P} \left(\frac{M}{P_{i,e} \frac{K-1}{P_c^4} + \xi} < \gamma_{th} \right) = \mathbb{P} \left(P_{i,e} > \left(\frac{M}{\gamma_{th}} - \xi \right) \frac{P_c^4}{K-1} \right)$$

$$= \begin{cases} 1 & \text{if } \gamma_{th} \geq \frac{M}{\xi} \\ \mathbb{P} \left(P_{i,e} > \left(\frac{M}{\gamma_{th}} - \xi \right) \frac{P_c^4}{K-1} \right) & \text{otherwise} \end{cases}.$$

When $\gamma_{th} < \frac{M}{\xi}$, from (5.33), we have

$$P_{out}(\gamma_{th}) \approx \left(\frac{b_e}{b_e + c_e} \right)^{2-K} \exp \left(- \frac{\left(\frac{M}{\gamma_{th}} - \xi \right) P_c^4}{(K-1)d_e(b_e + c_e)} \right)$$

$$- \frac{c_e}{b_e + c_e} \sum_{n=0}^{K-3} \frac{1}{\Gamma(n+1)} \left(\frac{b_e}{b_e + c_e} \right)^{n-K+2} \Gamma \left(n+1, \frac{\left(\frac{M}{\gamma_{th}} - \xi \right) P_c^4}{(K-1)d_e c_e} \right), \quad (5.34)$$

where $\Gamma(s, x) \triangleq \int_x^\infty t^{s-1} e^{-t} \mathrm{d}t$ is the upper incomplete gamma function [27]. This outage probability expression is too complex for useful insights. A simplified one is derived in the following proposition for systems with good CSI quality.

Proposition 5. *Define*

$$D \triangleq \frac{\left(2\left(1 - P_c\right) + \frac{1}{P} + \frac{K}{Q}\right) P_c^3}{(K-1) d_e (b_e + c_e)}.$$

When $E_t \gg 1$ and $M \gg \gamma_{th}\left(2 d_e c_e (1 + \frac{c_e}{b_e}) K(K-1) + \frac{1}{P} + \frac{K}{Q}\right)$, we have

$$P_{out}(\gamma_{th}) \approx \left(\frac{b_e}{b_e + c_e}\right)^{2-K} \exp\left(D - \frac{M P_c^4}{\gamma_{th}(K-1) d_e (b_e + c_e)}\right). \qquad (5.35)$$

Proof. By the definition of P_c and E_t in (5.1), when $E_t \gg 1$, we have $P_c \approx 1$. Thus $\xi \approx 1/P + K/Q$. Further define that

$$a \triangleq \frac{b_e \left(\frac{M}{\gamma_{th}} - \xi\right) P_c^4}{(K-1) d_e c_e (b_e + c_e)}.$$

Thus, when $M \gg \gamma_{th}\left(2 d_e c_e (1 + \frac{c_e}{b_e}) K(K-1) + \frac{1}{P} + \frac{K}{Q}\right)$, we have $a \gg 2K > 1$ and therefore

$$\frac{\left(\frac{M}{\gamma_{th}} - \xi\right) P_c^4}{(K-1) d_e c_e} \gg 1.$$

Then, from [27, 8.357] we know that

$$\Gamma\left(n+1, \frac{\left(\frac{M}{\gamma_{th}} - \xi\right) P_c^4}{(K-1) d_e c_e}\right) \approx \left(\frac{\left(\frac{M}{\gamma_{th}} - \xi\right) P_c^4}{(K-1) d_e c_e}\right)^n \exp\left(-\frac{\left(\frac{M}{\gamma_{th}} - \xi\right) P_c^4}{(K-1) d_e c_e}\right).$$

With this approximation, the outage probability expression in (5.34) can be reformulated as

$$P_{out}(\gamma_{th}) \approx \left(\frac{b_e}{b_e + c_e}\right)^{2-K} \exp\left(-\frac{\left(\frac{M}{\gamma_{th}} - \xi\right) P_c^4}{(K-1) d_e (b_e + c_e)}\right)\left[1 - \frac{\sum_{n=0}^{K-3} \frac{a^n}{\Gamma(n+1)}}{\frac{b_e + c_e}{c_e} \exp(a)}\right]$$

$$= \left(\frac{b_e}{b_e + c_e}\right)^{2-K} \exp\left(-\frac{\left(\frac{M}{\gamma_{th}} - \xi\right) P_c^4}{(K-1) d_e (b_e + c_e)}\right)\left[1 - \frac{c_e}{b_e + c_e} \frac{\sum_{n=0}^{K-3} \frac{a^n}{\Gamma(n+1)}}{\sum_{n=0}^{\infty} \frac{a^n}{\Gamma(n+1)}}\right]$$

Now, we examine the increasing rate of $\frac{a^n}{\Gamma(n+1)}$ with respect to n. First, we have

$$\frac{\mathrm{d}\frac{a^n}{\Gamma(n+1)}}{\mathrm{d}n} = \frac{a^n(\ln a + \gamma - \sum_{k=1}^{n} \frac{1}{k})}{\Gamma(n+1)},$$

80

where γ is the Euler$-$Mascheroni constant. Let a_I be the largest integer that is no larger than a. As $a_I \gg 1$, we have $\gamma \approx -\ln a_I + \sum_{k=1}^{a_I} 1/k$. Thus, when $n = a_I$, $\mathrm{d}\left(a^n/\Gamma(n+1)\right)/\mathrm{d}n \approx 0$. Therefore, $a^n/\Gamma(n+1)$ is increasing with n when $n < a_I$.

As $a_I > 2K$, we have

$$\frac{\sum_{n=0}^{K-3} \frac{a^n}{\Gamma(n+1)}}{\frac{b_e+c_e}{c_e}\exp(a)} = \frac{\sum_{n=0}^{K-3} \frac{a^n}{\Gamma(n+1)}}{\frac{b_e+c_e}{c_e}\sum_{n=0}^{\infty}\frac{a^n}{\Gamma(n+1)}} < \frac{\sum_{n=0}^{K-3} \frac{a^n}{\Gamma(n+1)}}{\sum_{n=0}^{K-3}\frac{a^n}{\Gamma(n+1)} + \sum_{n=K-2}^{2K-5}\frac{a^n}{\Gamma(n+1)}}.$$

Further, $\frac{a^{K-2}}{\Gamma(K-1)} = \frac{a}{K-2}\frac{a^{K-3}}{\Gamma(K-2)} > 2\frac{a^{K-3}}{\Gamma(K-2)}$, which means that the least term in the summation $\sum_{n=K-2}^{2K-5}\frac{a^n}{\Gamma(n+1)}$ is twice of the largest term in the summation $\sum_{n=0}^{K-3}\frac{a^n}{\Gamma(n+1)}$. Besides, there are $K-2$ terms in both summations. Comparing each pair of terms, the ratio is $a^{K-2}/\Gamma(K-1) \gg 1$ since $a \gg 2K$. Thus we can conclude that

$$\frac{c_e}{b_e+c_e}\frac{\sum_{n=0}^{K-3}\frac{1}{\Gamma(n+1)}(a)^n}{\sum_{n=0}^{\infty}\frac{1}{\Gamma(n+1)}(a)^n} \ll \frac{1}{3}.$$

Therefore, this term can be ignored and the approximation in (5.35) is obtained. $\qquad\square$

We can see that the outage probability approximation in (5.35) is tight when the number of relay antennas is much larger than the number of source-destination pairs and the training power and transmit powers are high. These conditions will result in a high received SINR. Thus, the approximation in (5.35) applies to the high SINR case.

Note that (5.35) can also be obtained by deleting the 2nd summation term in the PDF formula in (5.33) and then integrating with the approximated PDF. This is because that, for the high SINR case, the outage probability is determined by the SINR distribution in the small SINR region, which is equivalently the high interference power region, corresponding to the tail of the PDF of the interference power. It can be seen from the PDF formula in (5.33) that, the 1st term has a heavier tail, thus dominates the outage probability behaviour, and the second summation term can be ignored.

Now, we explore insights from (5.35). As b_e, c_e, d_e are irrelevant with P or Q, we can see that the outage probability scales as $\exp\left(\frac{P_c^3}{P(K-1)d_e(b_e+c_e)}\right)$ with P and scales as $\exp\left(\frac{KP_c^3}{Q(K-1)d_e(b_e+c_e)}\right)$ with Q. Firstly, it shows the natural phenomenon

81

that increasing P or Q will decrease the outage probability. Also, we can see that the outage probability curve with respect to Q has a sharper slope than that with P. For example, let $P = Q = \alpha$, doubling P alone will shrink the outage probability by a factor of $\exp\left(\frac{P_c^3}{2(K-1)d_e(b_e+c_e)\alpha}\right)$, while doubling Q alone will shrink the outage probability by a factor of $\exp\left(\frac{KP_c^3}{2(K-1)d_e(b_e+c_e)\alpha}\right)$, which is K powers of the shrinkage of the doubling-P case. Furthermore, increasing the user and relay transmit power will not make the outage probability diminish to zero. An error floor exists due to the user-interference. On the other hand, increasing the number of relay antennas to infinity leads to faster decrease in the outage probability and makes it approach zero.

Note that, in our analysis, we assume $M \gg 1$ but does not go to infinity, so terms with $1/\sqrt{M}$ are not treated as asymptotically small and thus are not ignored. If $M \to \infty$ and $P_c \to 1$, the $1/\sqrt{M}$ terms can be seen as 0 and we will have $P_{out}(\gamma_{th}) \approx \left(\frac{(K-1)\sqrt{3.5}}{\sqrt{M}}\right)^{2-K} \exp\left(-\frac{M}{2\gamma_{th}}\right)$. However, this asymptotic analysis is not practical, because the number of massive MIMO antennas is usually a few hundreds in practice, so that $\sqrt{M}$ may not be much larger than other parameters such as K, P, Q.

5.5.2   ABER analysis

ABER is anther important metric of communication performance. Due to the complexity of the SINR distribution, ABER analysis of the massive MIMO relay network is not available in the literature. For the linearly increasing SINR case, the ABER can be analyzed as below.

Denote the ABER as $P_b(e)$. It is given by

$$P_b(e) = \int_0^\infty P_b(e|r) f_{\text{SINR}}(r) \mathrm{d}r, \tag{5.36}$$

where $P_b(e|r)$ is the conditional error probability and $f_{\text{SINR}}(r)$ is the PDF of the SINR.

As is known, for channels with additive white Gaussian noise, $P_b(e|r)$ is in the form of $A\mathrm{erfc}\left(\sqrt{Br}\right)$, or equivalently $2AQ\left(\sqrt{2Br}\right)$ for several Gray bit-mapped constellations employed in practical systems. Here $\mathrm{erfc}(x)$ and $Q(x)$ are comple-

mentary error function and Gaussian Q-function respectively. A and B are constants depended on the modulation, for example, for BPSK, $A = 0.5$, $B = 1$.

On the other hand, with the PDF of the interference power and the SINR approximation in the linearly increasing SINR case, the PDF of the SINR can be derived as below.

$$
f_{\text{SINR}}(r) = \frac{(b_e + c_e)^{K-3} M P_c^4}{r^2 (K-1) d_e b_e^{K-2}} \exp\left(-\frac{\left(\frac{M}{r} - \xi\right) P_c^4}{(K-1) d_e (b_e + c_e)}\right)
$$
$$
- \sum_{n=0}^{K-3} \frac{(b_e + c_e)^{K-n-3} M P_c^{4n+4}}{\Gamma(n+1)((K-1)d_e)^{n+1} c_e^n b_e^{K-n-2}} \frac{\left(\frac{M}{r} - \xi\right)^n}{r^2} \times
$$
$$
\exp\left(-\frac{\left(\frac{M}{r} - \xi\right) P_c^4}{(K-1)d_e c_e}\right), r \in \left(0, \frac{M}{\xi}\right). \tag{5.37}
$$

With $P_b(e|r)$ in (5.36) and $f_{\text{SINR}}(r)$ in (5.37), an approximation on the ABER is derived in the following proposition.

Proposition 6. *When $E_t \gg 1$ and $M \gg 2 d_e c_e (1 + c_e/b_e) K(K-1) + \frac{1}{P} + \frac{K}{Q}$, the ABER can be approximated as*

$$
P_b(e) \approx A \left(\frac{b_e}{b_e + c_e}\right)^{2-K} \exp\left(D - 2P_c^2 \sqrt{\frac{BM}{(K-1)d_e(b_e + c_e)}}\right). \tag{5.38}
$$

Proof. The PDF of the SINR in (5.37) can be rewritten as

$$
f_{\text{SINR}}(r) = \frac{(b_e + c_e)^{K-3} M P_c^4}{r^2 (K-1) d_e b_e^{K-2}} \exp\left(-\frac{\left(\frac{M}{r} - \xi\right) P_c^4}{(K-1) d_e (b_e + c_e)}\right) \times
$$
$$
\left[1 - \frac{\sum_{n=0}^{K-3} \frac{\left(\frac{b_e\left(\frac{M}{r} - \xi\right) P_c^4}{(K-1)d_e c_e(b_e+c_e)}\right)^n}{\Gamma(n+1)}}{\exp\left(\frac{b_e\left(\frac{M}{r} - \xi\right) P_c^4}{(K-1)d_e c_e(b_e+c_e)}\right)} \right]. \tag{5.39}
$$

As the ABER is determined by the PDF when r is small [14], we can assume $r < 1$. With $E_t \gg 1$ and $M \gg 2 d_e c_e (1 + c_e/b_e) K(K-1) + \frac{1}{P} + \frac{K}{Q}$, similarly as the proof of Proposition 5, we can show that

$$
\frac{\sum_{n=0}^{K-3} \left(\frac{b_e\left(\frac{M}{r} - \xi\right) P_c^4}{(K-1)d_e c_e(b_e+c_e)}\right)^n}{\Gamma(n+1) \exp\left(\frac{b_e\left(\frac{M}{r} - \xi\right) P_c^4}{(K-1)d_e c_e(b_e+c_e)}\right)} \ll 1,
$$

83

and thus this term can be ignored.

With the approximated PDF, the ABER can be derived by solving

$$\int_{r=0}^{M/\xi} A\,\mathrm{erfc}(\sqrt{B}r) f_{\mathrm{SINR}}(r) dr.$$

As the ABER is determined by the region when r is small, we can replace the integration region with $\int_{r=0}^{\infty}$ for tractability. Then, by substituting $\mathrm{erfc}(x) = \Gamma(\frac{1}{2}, x^2)/\sqrt{\pi}$, using the integration formula $\int_{0}^{\infty} \exp(-\mu x)\Gamma(v, \frac{a}{x})\mathrm{d}x = 2a^{v/2}\mu^{v/2-1}K_v(2\sqrt{\mu a})$ [27], and using $K_{\frac{1}{2}}(x) = \sqrt{\frac{\pi}{2x}}\exp(-x)$ [95], the ABER approximation in (5.38) is obtained. □

We can see from (5.38) that increasing M will make the ABER decrease and approach zero. Besides, we can see from (5.38) that for very large M the ABER behaves as $C\exp\left(-C'\sqrt{M}\right)$. As is known, the ABER of traditional MIMO system with M transmit antennas and 1 receive antenna under Rayleigh fading is $C_m\mathrm{SINR}^{-C'_m M}$. This shows different ABER behaviour in massive MIMO systems, where the ABER decreases exponentially with respect to $\sqrt{M}$. If using the diversity gain definition of traditional MIMO system [29], the massive relay network will have infinite diversity gain.

Besides, comparing (5.38) with (5.35), we see that the ABER and the outage probability has the same scaling with P and Q respectively. Thus P, Q scaling analysis for the outage probability also applies to the ABER. In addition, if the threshold is set as $\gamma_{th} = \sqrt{\frac{MP_c^4}{4B(K-1)d_e(b_e+c_3)}}$, the ABER equals A times the outage probability. Thus, there is a easy transformation between the two metrics.

5.6 Simulation Results

In this section, simulation results are shown to verify the analytical results. In Fig. 5.1, the average SINR scaling with the number of relay antennas M is shown for the five cases in Table 5.1. In the table, $\lfloor\sqrt{M}\rfloor$ is the floor function that rounds $\sqrt{M}$ towards minus infinity. From the settings of P_c, P, Q, and K, we can find their scales with M and then the r_s values are calculated based on the SINR scaling law in (5.23). The first three cases are constant SINR cases. In Case 4 and Case 5, the

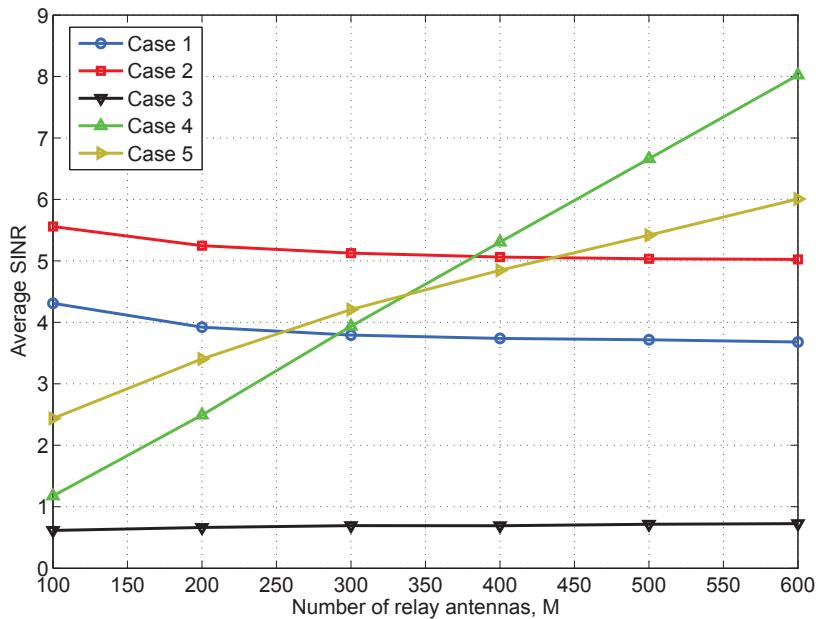|	P_c	P	Q	K	r_s
Case 1	0.8	10	10	$\frac{M}{10}$	0
Case 2	$\frac{100}{M}$	10	10	10	0
Case 3	0.8	10	$\frac{1}{\sqrt{M}}$	$\lfloor\sqrt{M}\rfloor$	0
Case 4	0.8	1	1	20	1
Case 5	$\frac{10}{\sqrt{M}}$	10	10	20	$\frac{1}{2}$

Table 5.1: Different cases for Fig. 5.1

Figure 5.1: Average SINR scaling for different number of relay antennas M for different scenarios.
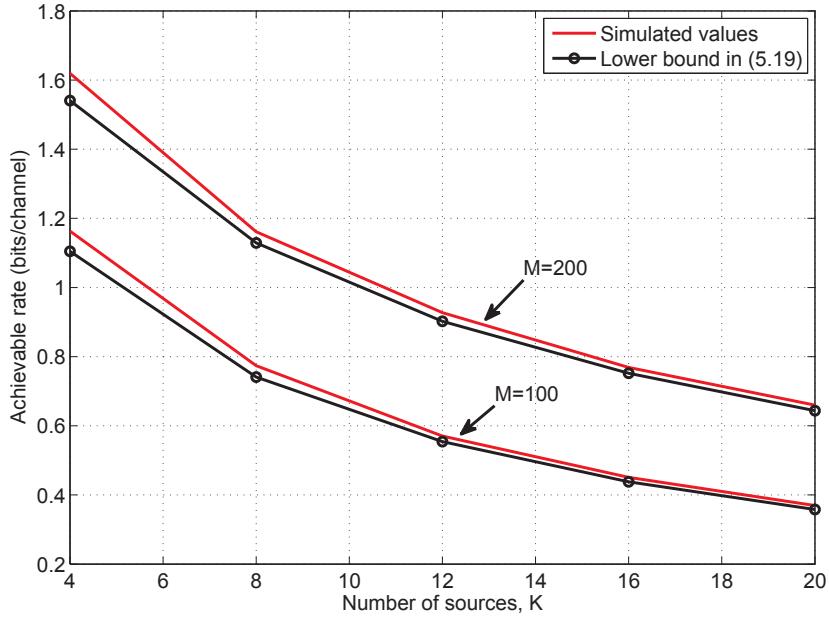
Figure 5.2: Achievable rate for different number of sources. $M = 200$ or 100, $P = Q = 0$ dB, $P_c = \frac{1}{2}$.

average SINR increases linearly with the relay antenna number, and the square root of the relay antenna number respectively. The figure verifies our scaling law results for different cases.

In Fig. 5.2, the average achievable rate per source-destination pair is simulated for different number of sources with 200 or 100 relay antennas. The source and the relay powers are set to be 0 dB. The CSI quality is set as $P_c = \frac{1}{2}$. From the figure, we can see that the approximated lower bound in (5.19) is very tight. With given number of relay antennas, the achievable rate per source-destination pair decreases as there are more pairs.

In Fig. 5.3, for a relay network with 20 or 10 source-destination pairs and 200 relay antennas, the simulated PDF of $P_{i,e}$ is shown. The CSI quality parameter is set as $P_c = 0.8$. The analytical expression in (5.33) is compared with the simulated values. We can see from Fig. 5.3 that the PDF approximation is tight for the whole parameter range. Especially, the approximation matches good at the tail when the interference power is large, which corresponds to the case with outage and transmission error. Thus the outage probability and ABER expressions derived with the

86
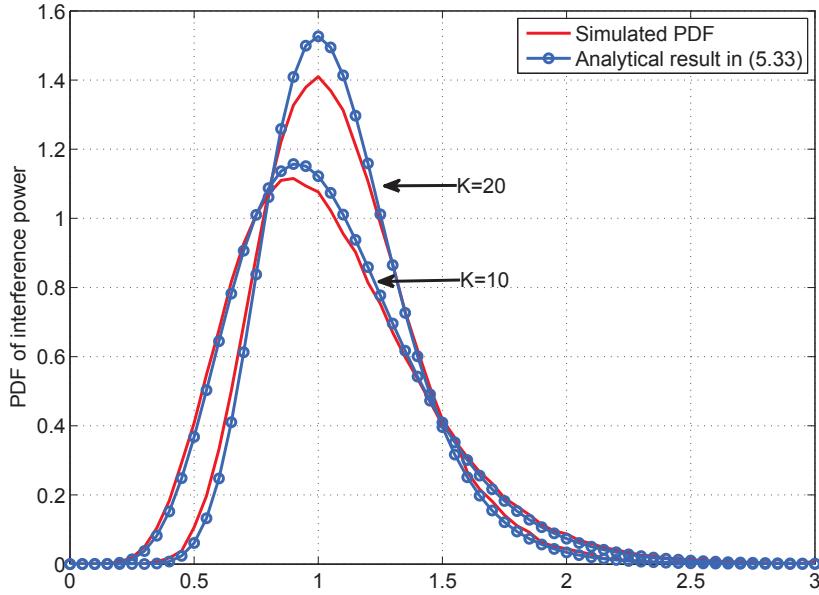
Figure 5.3: PDF of interference power. $K = 20$ or 10, $P_c = 0.8$, $M = 200$.

PDF approximation will be tight.

Fig. 5.4 and Fig. 5.5 are on the outage probability for different number of sources and different number of relay antennas. The analytical expressions in (5.34) and (5.35) are compared with the simulated values. The transmit powers of the users and the relay are set as 10 dB. The CSI quality parameter is $P_c = 0.95$. For Fig. 5.4, the number of relay antennas is 200, and the SINR threshold is 6 dB. For Fig. 5.5, the number of sources is 8 or 12 and the SINR threshold is 8 dB. We can see that our analytical result in (5.34) and the further approximation in (5.35) are both tight for all the simulated parameter ranges.

In Fig. 5.6 and Fig. 5.7, the ABER for BPSK is simulated. The analytical approximation in (5.38) is compared with the simulated values. Fig. 5.6 is for different number of sources with $M = 200$ or 300, $P = Q = 10$ dB and $P_c = 0.95$. Fig. 5.7 is for different number of relay antennas with $K = 8$ or 12, $P = Q = 10$ dB and $P_c = 0.95$. From the figures, we can see that the analytical results in (5.38) is tight for the simulated values.
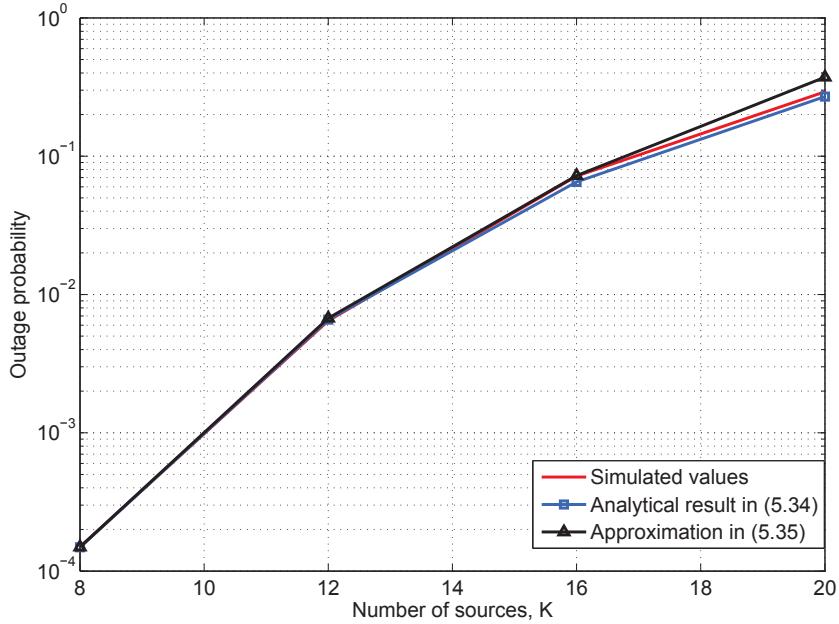
Figure 5.4: Outage probability for different number of sources. $M = 200, P = Q = 10$ dB, $\gamma_{th} = 6$ dB, and $P_c = 0.95$

5.7 Conclusion

In this chapter, we analyzed the performance of a massive MIMO relay network with multiple source-destination pairs under MRC/MRT relaying with imperfect CSI. Firstly, the performance scaling law is analyzed which shows that the scale of the sum-rate is decided by the summation of the scale of the CSI quality plus the smaller of the per-pair transmission power of the two hops. With this result, typical scenarios and trade-off between parameters are shown. Our scaling law is comprehensive as it takes into considerations many network parameters, including the number of relay antennas, the number of source-destination pairs, the source transmit power and the relay transmit power. Then, a sufficient condition for asymptotically deterministic SINR is derived, based on which new network scenarios for systems with the asymptotically deterministic property are found and trade-off between the parameters is analyzed. At last, we specify the necessary and sufficient condition for networks whose SINR increases linearly with the number of relay antennas. In addition, our work shows that for this case the interference power does
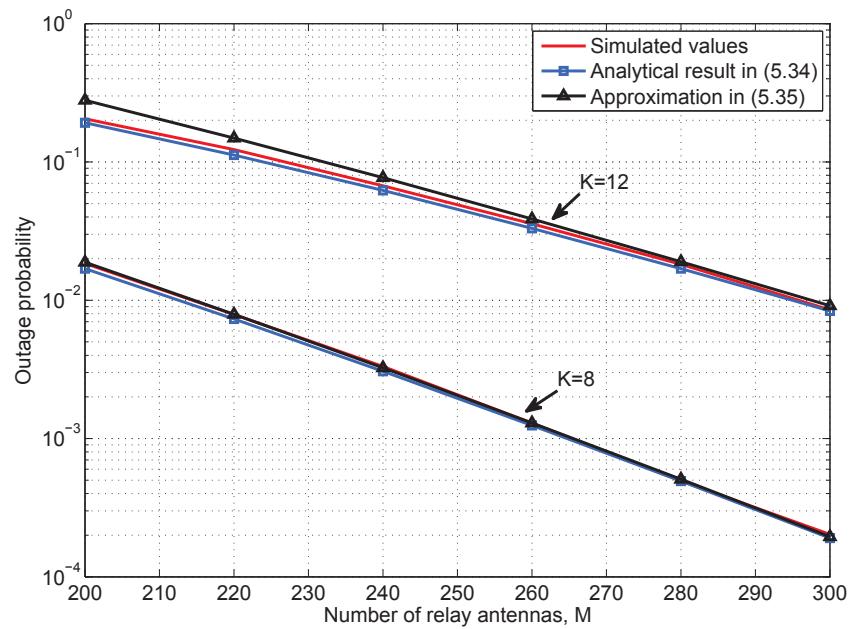
Figure 5.5: Outage probability for different number of relay antennas. $K = 8$ or $12, P = Q = 10$ dB, $\gamma_{th} = 8$ dB, $P_c = 0.95$.
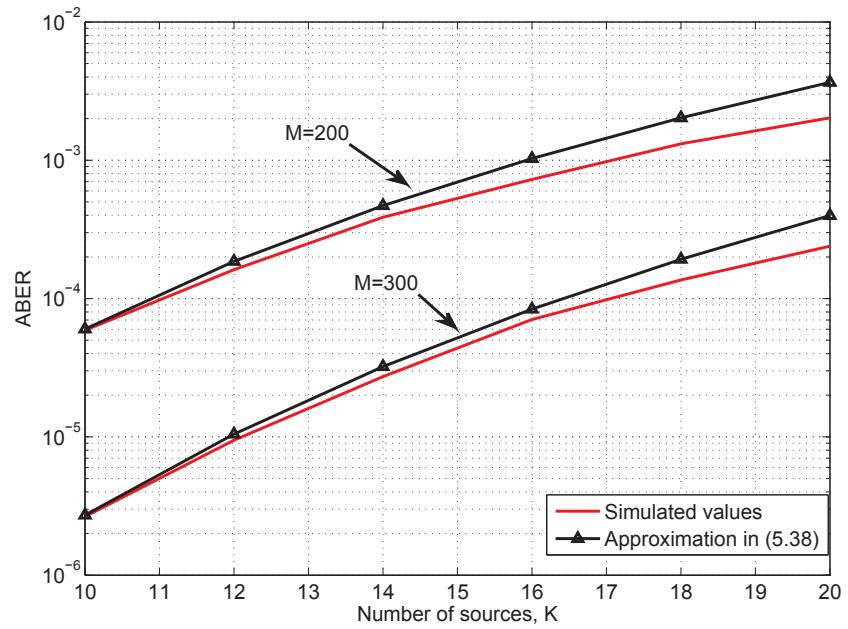
Figure 5.6: ABER of BPSK for different number of users K. $M = 200$ or $300, P = Q = 10$ dB, $P_c = 0.95$.
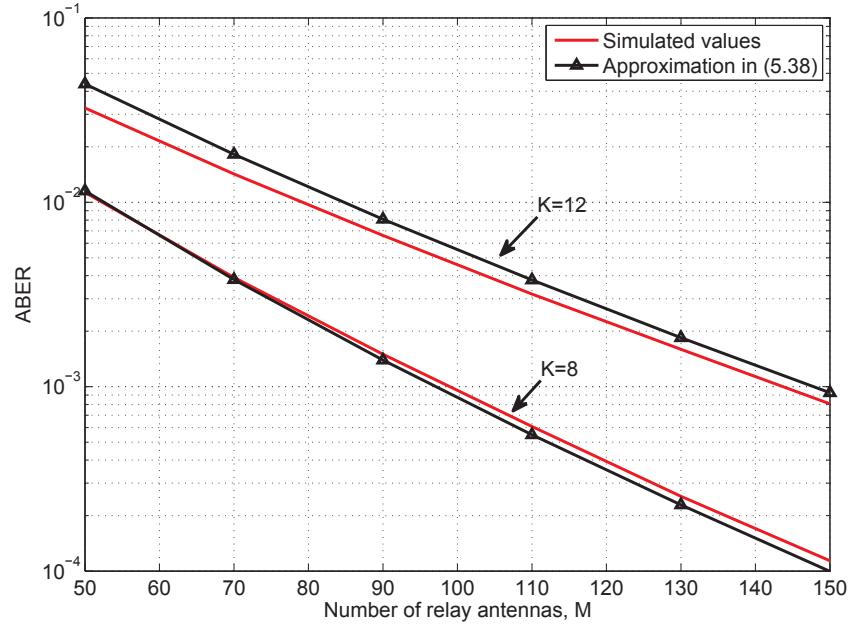
Figure 5.7: ABER of BPSK for different number of relay antennas M. $K = 8$ or 12, $P = Q = 10$ dB, $P_c = 0.95$.

not become asymptotically deterministic and derived the PDF of the interference power in closed-form. Then the outage probability and ABER expressions for the relay network are obtained and their behaviour with respect to network parameters are analysed. Simulations show that the analytical results are tight.

$\sim$

Chapter 6

SVD-Based Rank Detection for Reduced-Rank Channel Matrix

In practical large-scale relay systems, the channel matrices often have reduced rank. Reliable detection of the channel rank is essential in achieving the significant gain provided by the configuration. Existing work on the channel rank detection assume a static channel model, so the proposed methods only consider the noise distributions while the distributions of the channels are not considered. In this chapter, we employ a random channel model and propose three threshold-based rank detection methods which take into account the distributions of both the channels and the noises. In our first algorithm, following existing single-threshold rank detection scheme, we rigorously derive an analytical lower bound on the correct rank detection probability and propose a systematic threshold selection method by maximizing the lower bound. Then we propose two new rank detection methods which use multiple thresholds, where each threshold corresponds to one possible rank value. The thresholds are optimized based on the derived lower bounds on the rank detection probability for different channel rank values. The convergence and complexity of the proposed algorithms are analyzed. Simulation results on the correct rank detection probability of the proposed schemes are provided to show their advantages over existing schemes.

6.1 Introduction

As introduced in Chapter 2, the MIMO or relay channel usually has reduced rank, especially when the dimension is large. An important issue for the reduced-rank MIMO/relay system is the channel rank detection and channel matrix estimation, since most MIMO/relay transceiver techniques require channel state information (CSI) at the transmitter side and/or the receiver side for smart signal processing. For example, in the MIMO multiplexing transmissions, the transmitter should align multiple data-streams with the eigenspaces of the channel. To achieve this, precise estimates on the channel rank and channel matrix are needed. This chapter is concerned with the channel rank detection, which is an important part of channel estimation. Thus, in what follows, we explain related literature on channel estimation and rank detection.

Some existing work on MIMO channel estimation focus on full-rank MIMO channel matrices with independent or correlated entries [97–100]. Their channel estimation schemes are entry-based, where the unknown channel matrix is parametrized by its entries. However, when the channel matrix has reduced rank, the number of its entries is larger than its true dimension, which means that the channel entries are not independent. In this case, entry-based parametrization becomes inefficient. A more sensible approach is to use SVD-based channel estimations [19–24, 26, 101, 102]. It was shown in [19, 102] that the ML estimation of the reduced-rank MIMO channel with Gaussian noise is the truncated SVD method. In truncated SVD, if the channel rank is known to be r, the MIMO channel matrix is estimated from the SVD of the received signal-plus-noise matrix, by keeping the largest r singular values and their corresponding singular vectors. While the truncated SVD method traces back to 1930's [103], it was rediscovered in [19, 102] for MIMO channel estimation and further improved to reduce the MSE of the estimation [20–24]. One improved scheme is shrinkage-and-threshold SVD, where the truncated singular values are further shrunk to remove the noise effect.

In both truncated SVD and shrinkage-and-threshold SVD, the channel estimation accuracy largely depends on the correct truncation of the singular values, which

is the rank detection of the channel matrix. Thus correct rank detection can improve the channel estimation quality, which is crucial for advanced MIMO and relay techniques such as beamforming and PA among data-streams. Furthermore, for a MIMO channel matrix, the rank is the indicator of how many data streams can be spatially multiplexed on the channel, and the data streams are represented by the singular values and the corresponding singular vectors. Similarly, for a multi-user system with multiple antennas at the base station or relay station, the rank of the channel matrix from all users to the station determines how many users can be served by the station within the same time-frequency bandwidth. Thus, accurate channel rank detection is an important part of channel estimation and is essential for MIMO and relay systems.

Various rank detection methods are available in the literature [19–24, 26]. In [19], a minimum description length-based rank detection was used for MIMO channels. A threshold is calculated at each instance of the channel to minimize the minimum description length. The minimum description length-based detection aims at minimizing the MSE. It also requires a large number of samples to work. In [26], several rank detection methods were proposed for a real-valued channel matrix. In the proposed method, the singular values or functions of singular values are compared with a threshold for rank detection. Lower and upper bounds on the threshold selection were discussed. [20] considered both the MIMO channel rank detection and the shrinkage of the singular values for the channel estimation. The rank detection scheme is also threshold-based, where two threshold selection methods are proposed. In the first method, the threshold parameter and the shrinkage parameter are jointly optimized for each channel realization to minimize a Stein unbiased risk estimate (SURE) of the MSE of the MIMO channel estimation. In the second method, the threshold is calculated from the distribution of the largest singular value of the noise matrix. In [21–24], for asymptotic MIMO channels where both dimensions of the channel matrix approach infinity with a fixed ratio, simple closed-form thresholds were derived for threshold-based rank detection.

In this chapter, we propose a threshold optimization method for the traditional single-threshold-based rank detection scheme and two new multiple-threshold-based

rank detection schemes. For the threshold optimization of the traditional single-threshold scheme, we first derive a closed-form lower bound on the probability of correct rank detection based on the a-priori channel rank distribution, then the optimal threshold is decided via the maximization of the lower bound. For the two new multiple-threshold schemes, different thresholds are used for different possible rank values, and each threshold is derived by maximizing the lower bound on the probability of correct rank detection when a specific rank value is assumed. Properties (e.g., well-posedness, convergence, complexity) of the two new schemes are discussed. Simulation results on the probability of correct rank detection of the proposed schemes are shown and their advantages over existing schemes are discussed.

Our model, problem formulation, and methods differ from existing ones in the following major aspects. First, we assume a random channel matrix where the channel entries follow Rayleigh flat-fading and the distribution of the channel coefficients is taken into account in the threshold optimization. On the contrary, in all existing work, the channels are assumed to be static and the distribution of the channel matrix is not used in the rank detection designs [19–26]. Also, in our model, a general training length and unitary training matrix are considered, while most existing work (e.g., [19–24, 26]) apply to identity training matrix only, where the training length equals the number of transmit antennas. Finally, in this work, we use the probability of correct rank detection as the performance measure and optimization objective, while existing work (except [26]) targeted at minimizing the MSE of the MIMO channel estimation [19–24].

In what follows, we clarify the major difference of our channel rank detection and channel estimation problem to the reduced-rank filter design problem in [25, 105, 109–115]. The goal of channel estimation is to estimate the channel matrix itself given limited training time. In communications, channel estimation is usually required in the transceiver designs to optimize the communication performance such as outage probability and BER. For the filter design, the goal is to obtain a precise estimate of the signals by filtering the observations, and usually the MSE is used as the design criterion. Regardless of the channel rank, a reduced-rank filter

94

can be used to lower the computational complexity and required training length. Naturally, precise channel rank detection and channel estimation can be helpful in reduced-rank filter design, but it is not a necessary step in filter design. Also the optimal rank for the filter may not be the true rank of the channel matrix.

The rest of this chapter is organized as follows. In Section 6.2, the reduced-rank channel model, the truncated SVD-based channel estimation, and the rank detection problem are presented. In Section 6.3, for the traditional single-threshold-based rank detection, we derive a lower bound on the probability of correct rank detection and propose to optimize the threshold via the maximization of the lower bound. Two new rank detection algorithms based on multiple thresholds corresponding to different possible rank values are introduced in Section 6.4, as well as discussions on their properties. Simulation results on the probability of correct rank detection in Section 6.5. Section 6.6 contains the conclusions.

6.2 Reduced-Rank Channel Model and Rank Detection Problem

6.2.1 Reduced-Rank Channel Model

We consider the estimation of the source-to-relay channel in a multi-user large-scale relay network, where there are M single-antenna sources and one relay with N antennas. Note that, the channel of relay-to-destination link can be estimated using the same method if channel reciprocity is assumed. The source-to-relay channel can be seen as a virtual MIMO channel. The channels are assumed to be flat-fading and block-fading. Denote the $M \times N$ channel matrix as $\mathbf{H}$ with its (i, j)-th entry being the channel from the ith transmit antenna to the jth receive antenna. Define

$$L \triangleq \max\{M, N\}, \tag{6.1}$$

$$K \triangleq \min\{M, N\}. \tag{6.2}$$

Denote the rank of the channel matrix as r, i.e., $\mathrm{rank}(\mathbf{H}) = r$. If $r = K$, the channel has full rank. If $r < K$, the channel has reduced rank. For a reduced-rank channel, the number of degrees of freedom in the channel matrix is less than its

95

dimension. A rank analysis on typical propagation environments shows that MIMO channels often experience rank deficiency [82,116–119]. For example, in [116], the rank distribution of 8×8 MIMO channels under four scenarios, namely generalized typical urban, generalized bad urban, generalized hill terrain, and generalized rural area, are reported. For all four scenarios, the probability that the channel matrix has full rank is zero. Especially for the scenarios of generalized typical urban and generalized hill terrain, the rank of the 8×8 MIMO channel is always no higher than 4.

A typical reduced-rank channel model is the finite scatterers/dimensional channel [82, 119], where the number of (clusters of) scatterers is finite. The rank of the channel matrix is not only constrained by the number of transmit and receive antennas, but also constrained by the number of scatterers. In the finite scatterers model in [119], under the assumptions that both the transmit and receive elements are isotropic and uncoupled and the signal bandwidth is narrow compared with the overall channel bandwidth, the channel matrix can be described as the product of a full-rank steering matrix and a propagation matrix. The propagation matrix models independent fast fading, geometric attenuation, and shadow fading. Following the model in [119] and the rank factorization in [25, 116], we assume that the channel matrix has the following decomposition:

$$\mathbf{H} \triangleq \mathbf{AB}, \tag{6.3}$$

where $\mathbf{A}$ is an $M \times r$ full-rank matrix and $\mathbf{B}$ is a $r \times N$ full-rank rectangular unitary matrix. In this work, we focus on Rayleigh fading by assuming that entries of $\mathbf{A}$ are i.i.d. and follow CSCG distribution, with zero-mean and unit-variance, i.e., $a_{ij} \sim \mathcal{CN}(0, 1)$, where a_{ij} is the (i, j)-th entry of $\mathbf{A}$. It can thus be shown straightforwardly that each entry of H has CSCG distribution and $h_{i,j} \sim \mathcal{CN}(0, \|\mathbf{b}_j\|_F^2)$, where $\mathbf{b}_j$ is the j-th column of $\mathbf{B}$. It is noteworthy that the rank detection schemes proposed in this paper are not constrained to Rayleigh distribution and can be extended to more general channel fading models. The schemes are also not constrained to the unitary $\mathbf{B}$ case and can be extended to any general deterministic $\mathbf{B}$ matrix.

6.2.2 Training Model

To estimate the channel matrix, a training process is needed. Denote the length of the training period as T and average transmit power used for each training time slot as P. During the training period, the transmitter sends $\sqrt{PT/M}\mathbf{S}$, where $\mathbf{S}$ is the $T \times M$ pilot matrix. For the observability of the channel rank detection model with respect to all possible rank values, we assume that $T \geq M$, which guarantees that the number of independent equations in the training equation is no less than the number of independent unknown coefficients in the channel matrix. We further assume that $\mathbf{S}$ is unitary, i.e., $\mathbf{S}^*\mathbf{S} = I_M$, which means that the pilot vector sent from each transmit antenna is orthogonal to each other and has the same energy. Denote the $T \times N$ matrix received at the receiver as $\mathbf{Y}$. We have

$$\mathbf{Y} = \sqrt{\frac{PT}{M}}\mathbf{S}\mathbf{H} + \mathbf{W}, \tag{6.4}$$

where $\mathbf{W}$ is the $T \times N$ noise matrix. Entries of the noise matrix are assumed to be i.i.d. CSCG random variables with zero-mean and unit-variance. The pilot and the noise are assumed to be independent to the channel matrix, which applies to most practical systems.

6.2.3 SVD-Based Channel Estimation and Rank Detection Problem

The channel estimation problem is to estimate $\mathbf{H}$ from the observation $\mathbf{Y}$. To do this, we first transform the training equation in (6.4) to obtain a more direct relationship between the channel and the (transformed) received signal. Define

$$\tilde{\mathbf{Y}} \triangleq \sqrt{\frac{M}{PT}}\mathbf{S}^*\mathbf{Y}, \tag{6.5}$$

which is the $M \times N$ transformed received signal matrix. By left-multiplying both sides of (6.4) with $\sqrt{\frac{M}{PT}}\mathbf{S}^*$, we have

$$\tilde{\mathbf{Y}} = \mathbf{H} + \tilde{\mathbf{W}}, \tag{6.6}$$

where

$$\tilde{\mathbf{W}} \triangleq \sqrt{\frac{M}{PT}}\mathbf{S}^*\mathbf{W}. \tag{6.7}$$

Since $\mathbf{S}^*$ is an $M \times T$ unitary matrix, $\tilde{\mathbf{W}}$ is an $M \times N$ random matrix. Entries of $\tilde{\mathbf{W}}$ can be shown to be i.i.d. CSCG random variables with zero-mean and their variances are $M/(PT)$. From (6.6), $\tilde{\mathbf{Y}}$ is a noisy observation of the channel matrix $\mathbf{H}$ with white Gaussian noises.

Let $\tilde{\mathbf{Y}} = \mathbf{P}\mathrm{diag}\{\sigma_1, \cdots, \sigma_K\}\mathbf{Q}^*$ be the SVD of $\tilde{\mathbf{Y}}$, where $\mathbf{P}$ and $\mathbf{Q}$ are $M \times K$ and $K \times N$ unitary matrices and σ_i's are in non-increasing order, i.e., $\sigma_1 \geq \cdots \geq \sigma_K \geq 0$. If the rank of $\mathbf{H}$ is known to be r, the ML estimation of $\mathbf{H}$ has been proved to be the truncated SVD of $\tilde{\mathbf{Y}}$ given as follows [103]:

$$\hat{\mathbf{H}} = \mathbf{P}\mathrm{diag}\left\{\sigma_1, \cdots, \sigma_r, 0, \cdots, 0\right\}\mathbf{Q}^*. \tag{6.8}$$

In (6.8), an estimation with rank-r is obtained by keeping the subspaces with respect to the r strongest singular values of $\tilde{\mathbf{Y}}$. Subspaces with respect to the $K - r$ smallest singular values are seen as the noise effect and are ignored. This process guarantees that the estimator has the same rank with the real channel.

On the other hand, if the channel rank is unknown or the channel has full rank, an entry-based ML estimation can be obtained as [97]

$$\hat{\mathbf{H}}_{\mathrm{entry}} = \sqrt{\frac{M}{PT}}(\mathbf{S}^*\mathbf{S})^{-1}\mathbf{S}^*\mathbf{Y}. \tag{6.9}$$

The entry-based estimation leads to a full-rank matrix. When $r = K$, i.e., the channel has full rank, the SVD-based estimation and entry-based estimation are equivalent. When $r < K$, i.e., the channel has reduced rank, the entry-based estimation will contain subspaces due to the noise effect only, thus have worse performance.

Therefore, the rank detection is an essential problem for the SVD-based estimation. Wrong rank detection will lead to channel estimation error. In addition, it can degrade the performance of the communication. If the detected rank is smaller than r, some singular values and the corresponding singular spaces existing in the channel matrix may be detected as the noise effect only, and the subspaces will be lost in the estimated channel. On the other hand, if the detected rank is larger than r, some singular values and the corresponding singular spaces which do not exist in the channel matrix but appear in $\tilde{\mathbf{Y}}$ because of the noise disturbance, may be detected as part of the channel matrix. In the communication, information and

98

power will be allocated to such subspaces, which cause loss of information and wasting of power since the sub-spaces do not exist in the channel. Our problem of this chapter is to detect the channel rank from the received signal $\mathbf{Y}$, or the transformed received signal $\tilde{\mathbf{Y}}$. It is noteworthy that the rank detection does not require extra training since the same observations can be used for both rank detection and channel estimation.

6.3   Single-Threshold-Based Rank Detection

We can detect the rank of $\mathbf{H}$ from the singular values of $\tilde{\mathbf{Y}}$. Threshold-based algorithm appears to be a natural and common strategy [26], where the rank of $\mathbf{H}$ is detected as the number of singular values of $\tilde{\mathbf{Y}}$ that are larger than the threshold. With this scheme, singular values of $\tilde{\mathbf{Y}}$ that are smaller than the threshold are seen as the effect of the noise only; while singular values of $\tilde{\mathbf{Y}}$ that are larger than the threshold are seen as the effect of non-zero component of the channel matrix with small noise disturbance.

Let ϵ_{th} be the threshold. Recall that σ_i's are the singular values of $\tilde{\mathbf{Y}}$ in non-increasing order. The single-threshold-based rank detection scheme, denoted as RD1, can be represented as follows:

$$\text{RD1}: \quad r = \max_{\sigma_i \geq \epsilon_{\text{th}}} \{i\}. \tag{6.10}$$

If no singular value is larger than ϵ_{th}, i.e., $\sigma_1 < \epsilon_{th}$, the rank detection result is set to be 1 since the rank of a MIMO channel cannot be zero. The algorithm is given in Algorithm 2.

1: **for** $(i = K : 1)$ **do**
2:    **if** $\sigma_i \geq \epsilon_{\text{th}}$ **then**
3:       The rank of $\mathbf{H}$ is detected as i; break;
4: **if** $\epsilon_{\text{th}} > \sigma_1$ **then**
5:    The rank of $\mathbf{H}$ is detected as 1.

Algorithm 2: Rank detection algorithm with single threshold

This single-threshold-based rank detection idea is not new and was proposed and used in [19–24, 26]. But the major challenge of using this scheme for rank

detection lies in the selection of the threshold ϵ_{th}. Appropriate selection of the ϵ_{th} value is crucial to the detected result.

In this section, we first derive a lower bound on the probability of correct detection in Section 6.3.1, then propose a systematic method for the threshold selection based on maximizing the lower bound in Section 6.3.2, and finally discuss the difference of the proposed method with existing ones in Section 6.3.3.

6.3.1 Derivation of a Lower Bound on the Conditional Probability of Correct Rank Detection

To find a systematic way of optimizing the threshold, we first derive a lower bound on the probability of correct detection conditioned on an arbitrary rank value of the channel matrix. The lower bound takes into consideration the system dimensions (e.g., T, M, and N), training power P, as well as the distributions of the channel coefficients and the noises. It will be used in the threshold optimization in later sections.

To help presenting our results, we first introduce the following definitions. Define the $K \times K$ matrix $\boldsymbol{F}^{(1)}(\mu)$ and the $r \times r$ matrix $\boldsymbol{F}^{(r)}(\mu)$, whose (i,j)-th entries are:

$$\left[\boldsymbol{F}^{(1)}(\mu)\right]_{i,j} \triangleq \gamma(L - K + i + j - 1, \mu), \qquad (6.11)$$

$$\left[\boldsymbol{F}^{(r)}(\mu)\right]_{i,j} \triangleq \Gamma(M - r + i + j - 1, \mu), \qquad (6.12)$$

where $\gamma(k, u)$ and $\Gamma(k, u)$ are the lower and upper incomplete gamma functions [104], respectively.

The following proposition on the probability of correct rank detection under the condition that the rank of the MIMO channel matrix is r (for $1 \le r \le K$) is derived.

Proposition 7. *If the rank of* $\mathbf{H}$ *is* r, *the probability of correct rank detection of Algorithm 2 with threshold* ϵ_{th} *has the following lower bound:*

$$\phi_r(\epsilon_{\text{th}}) \triangleq C_1 C_2 \cdot \det\left(\boldsymbol{F}^{(r)}\left(4\epsilon_{\text{th}}^2\right)\right) \det\left(\boldsymbol{F}^{(1)}\left(\frac{PT}{M}\epsilon_{\text{th}}^2\right)\right), \qquad (6.13)$$

where

$$C_1 = \prod_{i=1}^{r} \left[(M - i)!(r - i)!\right]^{-1}, \quad C_2 = \prod_{i=1}^{K} \left[(L - i)!(K - i)!\right]^{-1}. \qquad (6.14)$$

100

Proof. Recall that σ_i's are the singular values of $\tilde{\mathbf{Y}}$ in non-increasing order, i.e., $\sigma_1 \geq \cdots \geq \sigma_r \geq \cdots \geq \sigma_K \geq 0$. Let λ_i and γ_i be the singular values of $\mathbf{H}$ and $\tilde{\mathbf{W}}$, respectively, both in non-increasing order, i.e, $\lambda_1 \geq \cdots \geq \lambda_r \geq 0$ and $\gamma_1 \geq \cdots \geq \gamma_r \geq \cdots \geq \gamma_K \geq 0$. Since $\mathrm{rank}(\mathbf{H}) = r$, we have $\lambda_{r+1} = \cdots = \lambda_K = 0$.

We will first show that when $\lambda_r \geq 2\epsilon_{\mathrm{th}} \geq 2\gamma_1$, our algorithms will detect the rank of $\mathbf{H}$ as r, which is the correct detection. According to [26, 120], from (6.6), we have for all $i = 1, \cdots, K$,

$$|\sigma_i - \lambda_i| \leq \gamma_i \leq \gamma_1. \tag{6.15}$$

By noticing that $\lambda_{r+1} = 0$, from (6.15) with $i = r + 1$, we have $\sigma_{r+1} \leq \gamma_1$. Thus when $\lambda_r \geq 2\epsilon_{\mathrm{th}} \geq 2\gamma_1$, we have $\sigma_{r+1} \leq \gamma_1 \leq \epsilon_{\mathrm{th}}$. Also from (6.15) with $i = r$, $\sigma_r \geq \lambda_r - \gamma_1 \geq 2\epsilon_{\mathrm{th}} - \gamma_1 \geq \epsilon_{\mathrm{th}}$. By noticing that σ_i's are in non-increasing order, we can conclude that the rank detection result of Algorithm 2 is r, which is the correct detection.

Thus, a lower bound on the probability of correct detection is obtained as follows,

$$\mathbb{P}(\text{correct detection}|\mathrm{rank}(\mathbf{H}) = r)$$
$$\geq \mathbb{P}\left(\lambda_r \geq 2\epsilon_{\mathrm{th}} \geq 2\gamma_1|\mathrm{rank}(\mathbf{H}) = r\right)$$
$$= \mathbb{P}\left(\lambda_r \geq 2\epsilon_{\mathrm{th}} \ \& \ \gamma_1 \leq \epsilon_{\mathrm{th}}|\mathrm{rank}(\mathbf{H}) = r\right)$$
$$= \mathbb{P}\left(\lambda_r \geq 2\epsilon_{\mathrm{th}}|\mathrm{rank}(\mathbf{H}) = r\right)\mathbb{P}(\gamma_1 \leq \epsilon_{\mathrm{th}}), \tag{6.16}$$

where the last step is because that γ_1, the largest eigenvalue of $\tilde{\mathbf{W}}$, is independent of both λ_r and the rank of $\mathbf{H}$.

Recall that our channel is modelled as $\mathbf{H} = \mathbf{AB}$, where the $M \times r$ matrix $\mathbf{A}$ has independent entries following $\mathcal{CN}(0, 1)$ and $\mathbf{B}$ is a $r \times N$ unitary matrix. Then $\mathbf{HH}^* = \mathbf{ABB}^*\mathbf{A}^{=}\mathbf{AA}^*$, which is an $M \times M$ central Wishart matrix with degree r. The singular values of $\mathbf{H}$ are the square roots of the eigenvalues of $\mathbf{HH}^*$. The CDF of the smallest non-zero eigenvalue of $\mathbf{HH}^*$ is known as follows [104]:

$$F_{\omega_r}(\mu) = 1 - C_1 \cdot \det(\boldsymbol{F}^{(r)}(\mu)), \tag{6.17}$$

where $\boldsymbol{F}^{(r)}$ and C_1 are defined in (6.12) and (6.14), respectively. Thus,

$$\mathbb{P}\left(\lambda_r \geq 2\epsilon_{\mathrm{th}}|\mathrm{rank}(\mathbf{H}) = r\right) = 1 - F_{\omega_r}(4\epsilon_{\mathrm{th}}^2) = C_1 \cdot \det(\boldsymbol{F}^{(r)}(4\epsilon_{\mathrm{th}}^2)). \quad (6.18)$$

Next, we calculate $\mathbb{P}(\gamma_1 \leq \epsilon_{\mathrm{th}})$. Recall that entries of $\tilde{\mathbf{W}}$ are i.i.d. following $\mathcal{CN}(0, M/PT)$. Thus, $(PT/M)\tilde{\mathbf{W}}\tilde{\mathbf{W}}^*$ is an $M \times M$ central Wishart matrix with degree N. The CDF of its largest eigenvalue is known to be [104]

$$F_{\omega_1}(\mu) = C_2 \cdot \det(\boldsymbol{F}^{(1)}(\mu)), \quad (6.19)$$

where $\boldsymbol{F}^{(1)}$ and C_2 are defined in (6.11) and (6.14), respectively. Therefore,

$$\mathbb{P}(\gamma_1 \leq \epsilon_{\mathrm{th}}) = C_2 \det\left(\boldsymbol{F}^{(1)}\left(\frac{PT}{M}\epsilon_{\mathrm{th}}^2\right)\right). \quad (6.20)$$

By using (6.18) and (6.20) in (6.16), the lower bound in (6.13) is obtained. $\quad\square$

6.3.2 Threshold Optimization

Assume that the a-priori probability mass function of the channel rank, $\mathbb{P}(\mathrm{rank}(\mathbf{H}) = r)$ for $r = 1, \cdots, K$, is known. Define

$$\phi(\epsilon_{\mathrm{th}}) \triangleq C_2 \det\left(\boldsymbol{F}^{(1)}\left(\frac{PT}{M}\epsilon_{\mathrm{th}}^2\right)\right) \cdot \sum_{r=1}^{K} C_1 \det\left(\boldsymbol{F}^{(r)}\left(4\epsilon_{\mathrm{th}}^2\right)\right) \mathbb{P}(\mathrm{rank}(\mathbf{H}) = r). \quad (6.21)$$

Given Algorithm 2 and threshold ϵ_{th}, the overall probability of correct rank detection can be lower bounded by $\phi(\epsilon_{\mathrm{th}})$. The derivations are as follows.

$$\begin{aligned}
&\mathbb{P}(\text{correct detection}) \\
&= \sum_{r=1}^{K} \mathbb{P}(\text{correct detection}|\mathrm{rank}(\mathbf{H}) = r)\mathbb{P}(\mathrm{rank}(\mathbf{H}) = r) \\
&\geq \sum_{r=1}^{K} \phi_r(\epsilon_{\mathrm{th}})\mathbb{P}(\mathrm{rank}(\mathbf{H}) = r) \quad (6.22) \\
&= C_2 \det\left(\boldsymbol{F}^{(1)}\left(\frac{PT}{M}\epsilon_{\mathrm{th}}^2\right)\right) \sum_{r=1}^{K} C_1 \det\left(\boldsymbol{F}^{(r)}\left(4\epsilon_{\mathrm{th}}^2\right)\right) \mathbb{P}(\mathrm{rank}(\mathbf{H}) = r) \quad (6.23)
\end{aligned}$$

where in (6.22) and (6.23), the results in Proposition 7 have been used.

We thus choose the threshold ϵ_{th} so that the lower bound is maximized, i.e.,

$$\epsilon_{\mathrm{th}}^* = \arg\max_{\epsilon_{\mathrm{th}}} \phi(\epsilon_{\mathrm{th}}). \quad (6.24)$$

The optimization problem in (6.24) is one-dimensional and can be optimally solved via exhaustive grid search. But there is a natural tradeoff between precision and computational complexity. For low computational complexity, in solving (6.24), we can find a zero point of $d \ln \phi(\epsilon_{\text{th}})/d\epsilon_{\text{th}}$ via bisection method and use it as the threshold. This low complexity method can result in sub-optimality when $d \ln \phi(\epsilon_{\text{th}})/d\epsilon_{\text{th}}$ has multiple zero-points.

In this section, for the traditional single-threshold-based rank detection, we rigorously derived an analytical lower bound on the correct rank detection probability, based on which a systematic threshold optimization scheme that maximizes this lower bound is proposed. The derived optimal threshold is adaptive to the number of transmit and receive antennas of the MIMO channel, the training length and power, as well as the distributions of the channel coefficients and the noise. Meanwhile, it is independent of the instantaneous channel values or singular values of the channel matrix. Thus the threshold optimization can be conducted off-line, which largely reduces the delay of channel rank detection in real applications.

It is noteworthy that the proposed method is not limited to Rayleigh fading channel model but can be extended to other distributions. Our lower bound calculation and threshold optimization need the distributions of the singular values of the channel matrix. For other channel fading models, even if no closed-form expressions for the distributions of the singular values are available, numerical estimations of the singular value distributions can be obtained via simulation, and our method can still be used. Especially as our methods can be conducted off-line, the computation complexity is not an issue in the real-time channel rank detection and channel estimation within a coherence interval of the channel.

Notice that in the proof of Proposition 7, we can loose our condition for the lower bound $\lambda_r \geq 2\epsilon_{\text{th}} \geq 2\gamma_1$ to $\lambda_r \geq 2\epsilon_{\text{th}} \geq 2\gamma_r$ without affecting the validity of the proof. With this change, another lower bounds on the conditional and overall probabilities of correct rank detection, denoted as $\tilde{\phi}_r(\epsilon_{\text{th}})$ and $\tilde{\phi}(\epsilon_{\text{th}})$, can be

obtained as follows:

$$\tilde{\phi}_r(\epsilon_{\text{th}}) \triangleq \mathbb{P}\left(\lambda_r \geq 2\epsilon_{\text{th}} | \text{rank}(\mathbf{H}) = r\right)\mathbb{P}(\gamma_r \leq \epsilon_{\text{th}}),$$

$$\tilde{\phi}(\epsilon_{\text{th}}) \triangleq \sum_{r=1}^{K} \tilde{\phi}_r(\epsilon_{\text{th}})\mathbb{P}(\text{rank}(\mathbf{H}) = r).$$

Since $\gamma_r < \gamma_1$, it can be shown straightforwardly that $\tilde{\phi}_r(\epsilon_{\text{th}}) > \phi_r(\epsilon_{\text{th}})$ and $\tilde{\phi}(\epsilon_{\text{th}}) > \phi(\epsilon_{\text{th}})$, which means that the new lower bounds are tighter, and we may obtain a better threshold by maximizing $\tilde{\phi}(\epsilon_{\text{th}})$. However, the expression for $\mathbb{P}(\gamma_r \leq \epsilon_{\text{th}})$ is much more complex than $\mathbb{P}(\gamma_1 \leq \epsilon_{\text{th}})$. Notice that it is the rth largest singular value of $\tilde{\mathbf{W}}$, not the smallest singular value since $\tilde{\mathbf{W}}$ has full rank with probability 1. Thus the calculation of the derivative of $\tilde{\phi}(\epsilon_{\text{th}})$ is more involved, and the maximization of $\tilde{\phi}(\epsilon_{\text{th}})$ has higher computational complexity. Our simulations show that, the use of $\tilde{\phi}(\epsilon_{\text{th}})$ improves the performance but the improvement is moderate. Thus, balancing the performance and computation complexity, we choose $\phi(\epsilon_{\text{th}})$ for our algorithms.

6.3.3 Difference to Existing Single-Threshold Schemes

In this subsection, we explain the difference of our work with existing results on threshold-based rank detection in [20, 22, 23, 26].

First, the research in [20, 22, 23, 26] are for real-valued channel matrix and real-valued noise matrix. Also, they consider the special case of $T = M$ and $\mathbf{S} = \mathbf{I}_m$. Our work applies for complex-valued channel matrix and complex-valued noise matrix, a general training length T where $T \geq M$, and unitary $T \times M$ pilot $\mathbf{S}$. In what follows, we explain existing threshold selections and calibrate the results to our model and notation.

In [26], lower and upper bounds on the threshold selection were provided. No specific threshold value or optimization method was given.

In [22] and [23], the channel rank detection and channel estimation problem were considered jointly for the asymptotic case, where the channel matrix dimensions approach to infinity but with a fixed ratio. The threshold for the rank detection was chosen to minimize the MSE of the truncated SVD channel estimation. The

rank detection follows the traditional single-threshold algorithm in Algorithm 2. After calibrating their results to our model, the threshold proposed in [23] is

$$\epsilon_{\text{th},[20]} = \sqrt{\frac{MN}{PT}} \sqrt{2\left(\frac{K}{L} + 1\right) + \frac{8K}{K + L + \sqrt{K^2 + 14KL + L^2}}},$$

and the threshold proposed in [22] is

$$\epsilon_{\text{th},[19]} = \left(1 + \sqrt{\frac{K}{L}}\right) \sqrt{\frac{MN}{PT}}.$$

In [20], the single-threshold Algorithm 2 was used and the threshold selection was based on the minimization of the SURE of the MSE of the channel estimation. Two methods were proposed. The first method needs numerical threshold optimization and the optimization problem changes with the instantaneous channel realization. Thus, the threshold optimization needs to be repeated for every coherence interval, impairing its practicality and efficiency. As in the second method, an analytical threshold was proposed. After calibrated to our model, the threshold is

$$\epsilon_{\text{th},[22]} = F_{\tilde{\mathbf{w}}_1}^{-1}\left(1 - \frac{1}{\sqrt{\log L}}\right),$$

where $F_{\tilde{\mathbf{w}}_1}^{-1}(x)$ represents the inverse function of the CDF of the largest singular value of the noise matrix $\tilde{\mathbf{W}}$ in (6.7).

In all three aforementioned works, the proposed threshold selection methods depend on the distribution of the noise matrix only, where [22,23] used the asymptotic behaviour of the singular values and singular vectors of the noisy observation matrix when the dimensions of the channel matrix approach infinity; and [20] used the distribution of the largest singular value of the noise matrix. It was assumed in their work that the channel matrix is deterministic and their results cannot take advantage of the distribution of the channel matrix. On the contrary, we adopt a random channel model and our threshold selection takes into account both the distribution of the channel coefficients and the distribution of the noises.

6.4 Multiple-Threshold Rank-Detection Methods

To use the rank detection scheme in Algorithm 2 with the proposed threshold in the previous section, the a-priori probabilities of the channel rank need to be known.

However, for some wireless communication systems, the channel rank distribution may not be precisely known due to the mobility and complexity of the signal propagation environment. In this case, a rank detection algorithm that does not rely on the channel rank distribution is required. In addition, the lower bound on the probability of correct rank detection in (6.21) provides the average rank detection performance over all possible rank values. It may not be sharp enough for one channel realization with a specific rank value. Thus, in this section, we propose two improved rank detection algorithms which do not need the rank distribution.

6.4.1 Rank Detection Algorithm with Multiple Thresholds

Instead of using only a single threshold for the rank detection as in Algorithm 2, we propose to use K thresholds $\epsilon_{\text{th},1}^*, \epsilon_{\text{th},2}^*, \cdots, \epsilon_{\text{th},K}^*$, each corresponding to one of the K possible rank values, $1, \cdots, K$. These thresholds are optimized by maximizing the lower bound on the probability of correct rank detection conditioned on the channel rank value, i.e.,

$$\epsilon_{\text{th},i}^* = \arg\max_{\epsilon} \phi_i(\epsilon), \tag{6.25}$$

where $\phi_i(\epsilon)$ is defined in (6.13).

$\epsilon_{\text{th},i}^*$ serves as the rank detection threshold when the channel rank is i, and $\epsilon_{\text{th},i+1}^*$ serves as the rank detection threshold when the channel rank is $i + 1$. Recall that $\sigma_1, \cdots, \sigma_K$ are ordered singular values of $\tilde{\mathbf{Y}}$. Thus, it is natural to detect the channel rank as i when both the following two conditions are satisfied C1) $\sigma_i \geq \epsilon_{\text{th},i}^*$ and C2) $\sigma_{i+1} < \epsilon_{\text{th},i+1}^*$. To help the presentation, we define the following set:

$$\mathcal{I} = \{i|\sigma_i \geq \epsilon_{\text{th},i}^* \,\&\, \sigma_{i+1} < \epsilon_{\text{th},i+1}^*, i = 1, \cdots, K - 1\},$$

which is the set of rank detection values that satisfy the two conditions. Since it is possible that $\mathcal{I}$ has 2 or more elements, for the uniqueness of the detection result, we detect the rank as the largest index that satisfies the two conditions. In other words, the detection rule, called RD2, can be represented as follows:

$$\text{RD2}: \quad r = \max_{i \in \mathcal{I}} \{i\}. \tag{6.26}$$

To guarantee the existence of a rank detection result, we also need to consider the case that there is no element in $\mathcal{I}$, i.e., $\mathcal{I} = \emptyset$. The following claim is proved.

Claim 1. *For two sequences of real numbers $a_1, \cdots, a_K$ and $b_1, \cdots, b_K$. If there exists no integer i (for $1 \leq i \leq K - 1$) such that $a_i \geq b_i$ and $a_{i+1} < b_{i+1}$, one of the following two cases must be true:*

1. *there exists an integer D, $1 \leq D \leq K$, such that $a_i < b_i$ for $i < D$, and $a_i \geq b_i$ for $i \geq D$*

2. *$a_i < b_i$ for all i.*

Proof. Assume that there exists no integer i (for $1 \leq i \leq K - 1$) such that $a_i \geq b_i$ and $a_{i+1} < b_{i+1}$. We consider the two cases $a_K < b_K$ and $a_K \geq b_K$ separately.

When $a_K < b_K$, if there exists an i such that $a_i \geq b_i$, let $i_{\max}$ be the largest i satisfying $a_i \geq b_i$. We have $a_{i_{\max}} \geq b_{i_{\max}}$ and $a_{i_{\max}+1} < b_{i_{\max}+1}$, which contradicts the assumption. Thus $a_i < b_i$ for all i.

When $a_K \geq b_K$, if there exists an i such that $a_i < b_i$, let $i_{\max}$ be the largest i satisfying $a_i < b_i$. Then, we will have $a_i < b_i$ for all $i \leq i_{\max}$, based on the same reasoning as above. In this case, $D = i_{\max} + 1$. When there does not exist an i such that $a_i < b_i$, meaning $a_i \geq b_i$ for all i, $D = 1$. $\qquad\square$

Based on Claim 1, when $\mathcal{I} = \emptyset$, we have either $\sigma_i < \epsilon^*_{\text{th},i}$ for $i < D$ and $\sigma_i \geq \epsilon^*_{\text{th},i}$ for $i \geq D$ ($1 \leq D \leq K$), in which case the rank detection result should be K; or $\sigma_i < \epsilon^*_{\text{th},i}$ for all $i = 1, \cdots, K$, in which case the rank detection result should be 1, which is the lowest possible rank for the random Rayleigh-fading channel matrix $\mathbf{H}$.

Given these discussions, our second rank detection scheme with multiple thresholds is described in Algorithm 3.

6.4.2 Iterative Rank Detection Algorithm with Multiple Thresholds

In our third rank detection algorithm, we iteratively use the K thresholds $\epsilon^*_{\text{th},1}, \cdots,$ $\epsilon^*_{\text{th},K}$ defined in (6.25) to refine our rank detection threshold and detection result. In each iteration, single threshold-based rank detection is performed and the threshold value is set using the rank detection result of the previous iteration. The iteration

```
1: while (i ≤ K − 1) do
2:     if σᵢ ≥ ϵ*_{th,i} and σ_{i+1} < ϵ*_{th,i+1} then
3:         The rank of H is detected as i; break;
4:     i = i + 1;
5: if (i == K) then
6:     if σ_K ≥ ϵ*_{th,K} then
7:         The rank of H is detected as K.
8:     if σ_K < ϵ*_{th,K} then
9:         The rank of H is detected as 1.
```

Algorithm 3: Rank detection algorithm with multiple thresholds.

ends when the rank detection result of the current iteration is the same as the result of the previous one. More specifically, first, a rank detection result r_1 is initialized, e.g., $r_1 = 1$, then $\epsilon^*_{\text{th},r_1}$ is used for the threshold of the next iteration. The new rank detection result is found using Algorithm 2, that is, the rank is detected as the maximum index i such that $\sigma_i \geq \epsilon^*_{\text{th},r_1}$ or 1 if all singular values of $\tilde{\mathbf{Y}}$ are smaller than $\epsilon^*_{\text{th},r_1}$. This new rank detection result is denoted as r_2. If $r_2 \neq r_1$, $\epsilon^*_{\text{th},r_2}$ is used as the threshold for the next iteration, and a new rank detection result can be obtained. The scheme is described in Algorithm 4. Please note that Lines 3-8 of Algorithm 4 are the same as Algorithm 2.

```
1: r₀ = 0, i = 1. Choose an integer value for r₁ between [1, K], e.g., r₁ = 1.
2: while rᵢ! = r_{i−1} do
3:     for m = K : 1 do
4:         if σₘ ≥ ϵ*_{th,rᵢ} then
5:             r_{i+1} = m; break;
6:     if ϵ*_{th,rᵢ} > σ₁ then
7:         r_{i+1} = 1.
8:     i = i + 1;
9: The rank of H is detected as rᵢ;
```

Algorithm 4: Iterative rank detection algorithm with multiple thresholds.

Claim 2. *If $\epsilon^*_{\text{th},1} \geq \epsilon^*_{\text{th},2} \geq \cdots \geq \epsilon^*_{\text{th},K}$, Algorithm 4 always converges.*

Proof. We prove the convergence by contradiction. Assume that the algorithm does not converge. From the algorithm, r_2 is the new rank detection result when using threshold $\epsilon^*_{\text{th},r_1}$. For any initial value for r_1, if $r_2 = r_1$, Algorithm 4 converges and

the rank detection result is r_1, which causes a contradiction. Next we consider the cases $r_2 > r_1$ and $r_2 < r_1$ separately.

Case 1: $r_1 < r_2$. Recall that r_2 is the new rank detection result when using threshold $\epsilon^*_{\text{th},r_1}$. Thus, we have either

- Case A: $r_2 = K$ and $\sigma_K \geq \epsilon^*_{\text{th},r_1}$; or

- Case B: $\sigma_{r_2} \geq \epsilon^*_{\text{th},r_1}$ and $\sigma_{r_2+1} < \epsilon^*_{\text{th},r_1}$.

Since $\epsilon^*_{\text{th},i}$'s are in non-increasing order, for Case A, we have $\sigma_K \geq \epsilon^*_{\text{th},r_1} \geq \epsilon^*_{\text{th},K}$. Thus the new rank detection result is K, i.e., $r_3 = K$. Then $r_3 = r_2$ and Algorithm 4 terminates, which contradicts the assumption. For Case B, since $r_1 < r_2$, we have $\sigma_{r_2} \geq \epsilon^*_{\text{th},r_1} \geq \epsilon^*_{\text{th},r_2}$. Thus the next rank detection result cannot be smaller than r_2, i.e., $r_2 \leq r_3$. If $r_2 = r_3$, Algorithm 4 terminates, which contradicts the assumption. Thus $r_2 < r_3$. The same situation happens for the next iterations. So, if Algorithm 4 does not converge, we will find an infinite strictly increasing integer sequence $r_1 < r_2 < r_3 < \cdots$. This contradicts the fact that r_m's are in the range of $[1, K]$.

Case 2: $r_1 > r_2$. Similarly, we have either Case A ($r_2 = 1$ and $\sigma_1 < \epsilon^*_{\text{th},r_1}$) or Case B listed above. For Case A, we have $\sigma_1 < \epsilon^*_{\text{th},r_1} \leq \epsilon^*_{\text{th},1}$. Thus the new rank detection result is 1, i.e., $r_3 = 1$ and Algorithm 4 terminates, which contradicts the assumption. For Case B, since $r_1 > r_2$, we have $\sigma_{r_2+1} < \epsilon^*_{\text{th},r_1} \leq \epsilon^*_{\text{th},r_2}$. Thus the next rank detection result cannot be larger than r_2, i.e., $r_3 \leq r_2$. If $r_2 = r_3$, Algorithm 4 terminates, which contradicts the assumption. Thus $r_2 < r_3$. The same situation happens for the next iterations. So, if Algorithm 4 does not converge, we will find an infinite strictly decreasing integer sequence $r_1 > r_2 > r_3 > \cdots$. This contradicts the fact that r_m's are in the range of $[1, K]$. $\qquad\square$

Claim 2 shows that when the thresholds corresponding to the K rank values are in non-increasing order, Algorithm 4 is guaranteed to converge. Also, from the proof of Claim 2, we can see that the convergence is guaranteed within K iterations. Our limited simulation results indicate that the algorithm converges very fast (within 2-3 iterations). Intuitively, as $\epsilon^*_{\text{th},i}$ is the threshold when the channel rank is i and the singular values are non-increasingly ordered, it is natural to have

$\epsilon_{\text{th},1}^*, \epsilon_{\text{th},2}^*, \cdots, \epsilon_{\text{th},K}^*$ in non-increasing order. However, we cannot prove this analytically. When violation of the ordering happens on one threshold, we can simply reset the threshold to be the average of the one before and the one after to fix the ordering problem and have guaranteed convergence.[1]

Although the convergence of Algorithm 4 is guaranteed, we cannot not guarantee the uniqueness of the rank detection solution with respect to different initial values for r_1. In other words, for different initial rank values, the algorithm may converge to different solutions. An example is as follows. Assume that $\sigma_1 > \epsilon_{\text{th},1}^* > \sigma_2 > \epsilon_{\text{th},2}^* > \cdots > \sigma_K > \epsilon_{\text{th},K}^*$. Then the final rank detection result of Algorithm 4 will equal to r_1 for any initial r_1-value.

6.4.3 Discussion on Complexity

The computational complexity of Algorithms 3 and 4 is composed of two parts: the calculations of $\epsilon_{\text{th},1}^*, \cdots, \epsilon_{\text{th},K}^*$ and the rank detection part.

The optimization of $\epsilon_{\text{th},i}^*$ only depends on the dimensions of the channel matrix M and N, the training time T, and the training power P. It is independent of the channel realization of each coherence interval. Thus the optimization can be conducted off-line. Further, the following lemma is proved which can be used to reduce the computational complexity of the optimization.

Lemma 5. *The function $\phi_r(\epsilon_{\text{th}})$ in (6.13) is a log-concave function of ϵ_{th}.*

Proof. For the Hermitian matrices $\boldsymbol{F}^{(1)}(\mu)$ and $\boldsymbol{F}^{(r)}(\mu)$, we can show that all their leading principal minors are positive when $\mu > 0$ from the definitions in (6.11), (6.12), and the CDFs in (6.17), (6.19). Thus the two matrices are positive definite and $\det\left(\boldsymbol{F}^{(1)}\left(\frac{PT}{M}\epsilon_{\text{th}}^2\right)\right)$ and $\det\left(\boldsymbol{F}^{(r)}\left(4\epsilon_{\text{th}}^2\right)\right)$ are log-concave functions since the determinant of a positive definite matrix is log-concave [81]. Based on [81], the product of log-concave functions is also log-concave. This ends the proof. $\square$

Notice that $\epsilon_{\text{th},r}^*$ is the maximum point of $\ln \phi_r(\epsilon_{\text{th}})$. With the log-concavity of $\phi_r(\epsilon_{\text{th}})$, we can find $\epsilon_{\text{th},r}^*$ by finding the unique zero-point of $d\ln\phi_r(\epsilon_{\text{th}})/d\epsilon_{\text{th}}$, using

---

[1]In simulations, the thresholds we numerically obtain from (6.25) are in non-increasing order, except when the channel dimension gets large. When the values of M and N are large, violation of the non-increasing order occasionally happens due to the limited precision of computer calculation.

bisection method. The calculations are as follows. From (6.13) and the definitions in (6.11) and (6.12), we have

$$\ln \phi_r(\epsilon_{\text{th}}) = \ln C_1 + \ln C_2 + \ln \det \left(\boldsymbol{F}^{(r)} \left(4\epsilon_{\text{th}}^2 \right) \right) + \ln \det \left(\boldsymbol{F}^{(1)} \left(\frac{PT}{M} \epsilon_{\text{th}}^2 \right) \right),$$

and

$$\frac{d(\ln \phi_r(\epsilon_{\text{th}}))}{d\epsilon_{\text{th}}} = \text{tr} \left[\left(\boldsymbol{F}^{(r)} \left(4\epsilon_{\text{th}}^2 \right) \right)^{-1} \mathbf{D}_r \right] + \text{tr} \left[\left(\boldsymbol{F}^{(1)} \left(\frac{PT}{M} \epsilon_{\text{th}}^2 \right) \right)^{-1} \mathbf{D}_1 \right], \quad (6.27)$$

where $\mathbf{D}_r \triangleq d \left(\boldsymbol{F}^{(r)} \left(4\epsilon_{\text{th}}^2 \right) \right) / d\epsilon_{\text{th}}$ and $\mathbf{D}_1 \triangleq d \left(\boldsymbol{F}^{(1)} \left(\frac{PT}{M} \epsilon_{\text{th}}^2 \right) \right) / d\epsilon_{\text{th}}$. The (i, j)-th entries of $\mathbf{D}_r$ and $\mathbf{D}_1$ are respectively

$$[\mathbf{D}_r]_{i,j} = -8\epsilon_{\text{th}} \exp \left(-4\epsilon_{\text{th}}^2 \right) \left(4\epsilon_{\text{th}}^2 \right)^{M-r+i+j-2}, \quad (6.28)$$

$$[\mathbf{D}_1]_{i,j} = 2\frac{PT}{M}\epsilon_{\text{th}} \exp \left(-\frac{PT}{M} \epsilon_{\text{th}}^2 \right) \left(\frac{PT}{M} \epsilon_{\text{th}}^2 \right)^{i+j-2}.$$

Next, we analyze the complexity of the rank detection part. For Algorithm 3, the total number of comparisons in the worst scenario is $2K$, thus the complexity is $\mathcal{O}(K)$. For Algorithm 4, in the worst case, the number of iterations is K; and for each iteration, at most $K + 1$ comparisons are needed. The overall number of comparisons is $K(K + 1)$. Thus the complexity is $\mathcal{O}(K^2)$. Notice that $K = \min\{M, N\}$. For large-scale relay systems with N relay antennas and M single-antenna users, where N is large (e.g., hundreds), the complexities of the two proposed rank detection algorithms are linear and quadratic in the number of users, respectively.

6.5 Simulation Results

In this section, simulation results are shown for Algorithm 2 with our proposed threshold optimization in Section 6.3, and the two new rank detection algorithms with multiple thresholds, Algorithm 3 and Algorithm 4, proposed in Section 6.4. We simulate the probability of correct rank detection for different average training power P. In our simulation, channel coefficients are generated as Rayleigh fading following the model in (6.3). $\mathbf{S}$ is generated as a random $T \times M$ unitary matrix following the isotropic distribution. For comparison, we also show the rank detection
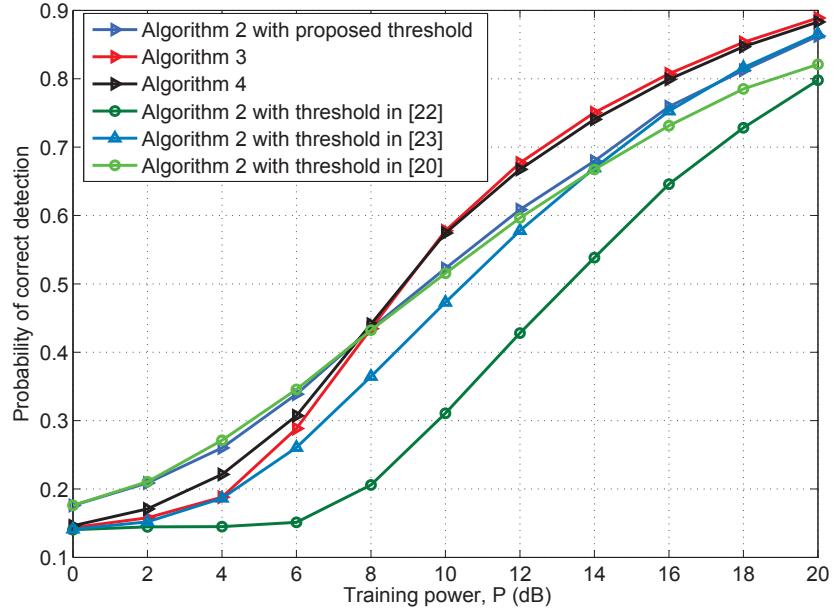
Figure 6.1: Probability of correct rank detection of 7×50 large-scale system for different average training power, with uniformly distributed channel rank values.

accuracy of Algorithm 2 with the threshold values proposed in [20, 22, 23]. While, the results in [20, 22, 23] are for $T = M$ and $\mathbf{S} = \mathbf{I}_M$, we extend them for a general $T \geq M$ and unitary $\mathbf{S}$ as explained in Section 6.3.3.

In Fig. 6.1, we consider the source-to-relay link of the large-scale relay system. The receiver is a relay station where 50 antennas are deployed. The transmitters are 7 single-antenna users or one user with 7 antennas. The training length is $T = 20$. The channel rank is assumed to be uniformly distributed over 1 to 7. The probabilities of correct rank detection for different average training powers are shown. We can observe from this figure that for the single-threshold algorithm, Algorithm 2, the proposed threshold has about the same performance as the threshold in [20] for the whole training power range and are much better than the ones in [22, 23]. The proposed multiple-threshold algorithms, Algorithm 3 and Algorithm 4, achieve considerably higher detection rate than Algorithm 2 when the training power is higher than 8 dB. Algorithm 3 is slightly better than Algorithm 4 at high training power, but is sightly worse at low training power.

6.6 Conclusion

In this chapter, we proposed novel threshold-based rank detection algorithms for reduced-rank large-scale relay systems. Different from previous work, we consider a system with random channel matrix model, a general training length, and unitary training matrix. Lower bounds on the probability of correct rank detection were derived using the distribution of the channel matrix and noise matrix, based on which the rank detection thresholds were optimized. In addition to the traditional single-threshold detection algorithm, we further proposed two low-complexity multiple-threshold algorithms. Compared with the existing schemes, our proposed schemes can achieve higher rank detection rate for various scenarios.

$$\sim$$

Chapter 7

Conclusions and Future Work

7.1 Conclusions

To meet the requirement of high throughput as well as wide coverage of future wireless networks, the novel infrastructure, large-scale relay network, which inherits the advantages of both relaying and massive MIMO, is investigated in this thesis. For this network, we analyzed the performance of popular relaying schemes. The derived neat closed-form expressions help us better understand the network performance with respect to different network parameters. We also designed novel rank detection algorithms that improves the rank detection accuracy for reduced-rank channel estimation. Specific contributions are as follows.

- Chapter 3 characterizes the performance of BRS in large-scale relay networks. Closed-form approximations for the average SNR and ergodic capacity are derived. The closed-form expressions will help to design the network parameters to satisfy certain requirements. Besides, from the average SNR, we conclude that the array gain of BRS is linear in the logarithm of the relay antenna number. Simulation results show that our results are tight for a wide range of the transmit power and relay number, and are significantly superior to existing closed-form results.

- Chapter 4 first presents solutions to the PA problem of distributed relay beamforming in a multi-user multi-relay network. An optimal solution as well as an approximate closed-form solution are proposed. Then, the user SNR and

the network sum-rate are analyzed when the relay number is large. The analysis shows that the received SNR of each user is linear with the number of relay antennas and also the transmit power, thus we can make the transmit power inversely proportional to the number of relay antennas while maintaining the same SNR performance. It also shows that the sum rate is independent of the instantaneous channel coefficients, and only depends on channel variances, thus the PA problem can be conducted off-line and its complexity is no longer a certain due to the large number of antennas.

- Chapter 5 is on the performance of MRC/MRT relaying in a multi-user massive MIMO relay network with imperfect CSI. Firstly, the performance scaling law is analyzed with all parameters, including the number of sources, the CSI quality, and the transmit powers of sources and relay, scaling with the number of relay antennas. It is shown that the scale of the sum-rate equals the summation of the scales of the CSI quality and the minimum of the per-pair transmission power of the two hops. From the scaling law, the trade-off between parameters and their effects on the performance are revealed. Then, for the asymptotically deterministic SINR case, a sufficient condition for this case is derived. The condition covers existing results as special cases. At last, for linearly increasing SINR case, the sufficient and necessary condition for this case is shown be that the number of source-destination pairs, the transmit powers of sources and relay and the CSI quality should all remain constant. In this case, we show that the interference power does not diminish and dominates the statistical properties of the SINR. With its PDF derived in closed-form, the outage probability and ABER expressions for the network are obtained. Their performance with respect to network parameters are analyzed.

- Chapter 6 investigates the threshold-based rank detection algorithms for reduced-rank large-scale relay networks. Lower bounds on the probability of correct rank detection are derived. Different from existing work, in the derivation the distribution of both channel matrix and noise matrix are considered, making

the lower bound more accurate. For the traditional single-threshold rank detection, the proposed algorithm provides an optimal threshold that maximizes the lower bound. Besides, the optimization can be conducted off-line as it is independent of the instantaneous channel states. In addition to the single-threshold algorithm, two novel multiple-threshold algorithms are proposed to improve the detection accuracy. Both algorithms are proved to be well-posed and their complexities are irrelevant with the number of relay antennas and only in the order of the number of users. Simulation results show that the proposed algorithms achieve higher correct rank detection rate that existing ones.

7.2   Future Work

In the following, we outline a couple of possible future research directions.

- Further analysis on the asymptotic performance of distributed relaying schemes with estimation error or quantized CSI.

  In this thesis, we assumed perfect CSI in analyzing the distributed relaying schemes. But, in practice, the channel estimation cannot be perfect. The estimation error will affect the performance of the relaying schemes. Thus, one possible extension of this thesis is to further analyze the effect of error on the performance of distributed relaying schemes in large-scale relay networks. We may also explore the conditions for the effect of estimation error to diminish asymptotically by using a large number of relay antennas.

  On the other hand, as stated in Chapter 2, distributed relaying schemes require communication between relay antennas to share some channel or signal information. To reduce this overhead, the shared information is often quantized. The quantization loss in the accuracy of the shared information will affect the performance. Future work of this thesis may analyze the effect of quantization on the performance of large-scale relay networks, and find the optimal quantization solution to meet certain performance requirements.

- Investigation on novel transmission schemes for large-scale relay networks. The large scale brings new challenges as well as opportunities for the design of transmission schemes. In this thesis, we proposed novel rank detection algorithms in Chapter 6 for the channel estimation of large-scale relay networks. In the future work, we may investigate other transmission designs to improve the performance of the network.

One possible direction is the design of the hybrid processor. Traditional relaying schemes require each antenna to freely receive and send signals with arbitrary phase and amplitude change. Thus, each antenna should connect to an individual radio-frequency (RF) chain. A RF chain contains amplifiers, mixers and analog-to-digital converters/digital-to-analog converters. While, for large-scale relay networks with centralized relay systems, where all antennas and RF chains are implemented on one relay station, it is often expensive to provide each antenna with a RF chain, because of its high cost and power consumption [10]. A practical solution is to use less RF chains than antennas elements. This design has been proposed for massive MIMO networks [123–130], where a hybrid processor is implemented by cascading a high-dimension RF beamformer and a low-dimension baseband processor. The analyses in those work show that the hybrid processor can achieve the promised gains of massive MIMO systems with less RF chains. Recently, this hybrid processor has been applied to single-user large-scale relay networks in [131]. But for a more practical multiple-user large-scale relay network, the hybrid processor still needs to be designed and analyzed.

Bibliography

[1] (2016, Feb.) Cisco visual networking index: Global Mobile Data Traffic Forecast Update, 20152020 White Paper. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html.

[2] M. Ghosh. (2014, Jun.) 22 interesting insights on global mobile usage: Ericsson. [Online]. Available: http://trak.in/tags/business/2014/06/04/22-interesting-insights-global-mobile-usage/.

[3] "IEEE Standard for Information Technology-Telecommunications and information exchange between systems-Local and metropolitan area networks-Specific requirements-Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band," *IEEE Std 802.11ad-2012*, pp. 1-628, Dec. 2012.

[4] "Technical Specication Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); LTE physical layer; General description (Release 11)," *3GPP 3GPP TS 36.201 V11.0.0 Std*, Oct. 2012.

[5] Q. Li, R. Hu, Y. Qian, and G. Wu, "Cooperative communications for wireless networks: techniques and applications in LTE-advanced systems," *IEEE Wireless Commun.*, vol. 19, no. 2, pp. 2229, Apr. 2012.

[6] S. Mohammad Razavizadeh, M. Ahn, and I. Lee, "Three-dimensional beamforming: A new enabling technology for 5G wireless networks," *IEEE Sig. Process. Mag.*, vol. 31, no. 6, pp. 94101, Nov. 2014.

[7] S. Chen and J. Zhao, "The requirements, challenges, and technologies for 5G of terrestrial mobile telecommunication", *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 36–43, May 2014.

[8] K. R. Liu, *Cooperative communications and networking*, Cambridge university press, 2009.

[9] M. Hasan, E. Hossain and D. I. Kim, "Resource allocation under channel uncertainties for relay-aided device-to-device communication underlaying LTE-A cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2322–2338, Apr. 2014.

[10] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Sig. Process. Mag.*, vol. 30, no. 1, pp. 40–60, 2013.

[11] Y. Jing and H. Jafarkhani, "Single and multiple relay selection schemes and their achievable diversity orders," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1414–1423, Mar. 2009.

[12] E. Koyuncu, Y. Jing, and H. Jafarkhani, "Distributed beamforming in wireless relay networks with quantized feedback," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1429–1439, Oct. 2008.

[13] Y. Zhao, R. Adve, and T. J. Lim, "Symbol error rate of selection amplify-and-forward relay systems," *IEEE Commun. Letter*, vol. 10, no. 11, pp. 757–759, Nov. 2006.

[14] A. Ribeiro, X. Cai, and G. B. Giannakis, "Symbol error probabilities for general cooperative links," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1264–1273, May 2005.

[15] S. S. Soliman and N. C. Beaulieu, "Dual-hop AF systems with maximum end-to-end SNR relay selection over Nakagami-m and Rician fading links," in *Proc. Int. Conf. Computing, Networking and Commun.*, San Diego, CA, Jan. 2013, pp. 155–161.

[16] S. I. Hussain, M. S. Alouini, and M. O. Hasna, "Performance analysis of best relay selection scheme for amplify-and-forward cooperative networks in identical Nakagami-m channels," in *Proc. IEEE Sig. Process. Advances in Wireless Commun.*, Jun. 2010, pp. 1–5.

[17] S. S. Ikki and M. H. Ahmed, "Performance analysis of generalized selection combining for amplify-and-forward cooperative-diversity networks," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2009, pp. 1–6.

[18] M. Torabi, W. Ajib, and D. Haccoun, "Performance analysis of amplify-and-forward cooperative networks with relay selection over Rayleigh fading channels," in *Proc. IEEE Veh. Technol. Conf.*, Barcelona, Spain, Apr. 2009, pp. 1–5.

[19] M. Nicoli and U. Spagnolini, "Reduced-rank channel estimation for time-slotted mobile communication systems," *IEEE Trans. Sig. Process.*, vol. 53, no. 3, pp. 926–944, Mar. 2005.

[20] J. Josse and S. Sardy, "Selecting thresholding and shrinking parameters with generalized SURE for low rank matrix estimation," *arXiv preprint arXiv:1310.6602*, 2013.

[21] A. A. Shabalin and A. B. Nobel, "Reconstruction of a low-rank matrix in the presence of Gaussian noise," *J. Multivariate Anal.*, vol. 118, pp. 67–76, 2013.

[22] M. Gavish and D. L. Donoho, "Optimal shrinkage of singular values," *arXiv preprint arXiv:1405.7511*, 2014.

[23] D. L. Donoho and M. Gavish, "The optimal hard threshold for singular values is 4/sqrt (3)," *IEEE Trans. Info. Theory*, vol. 60, no. 8, pp. 5040-5053, Aug. 2014.

[24] R. Nadakuditi, "Optshrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage," *IEEE Trans. Info. Theory*, vol. 60, no. 5, pp. 3002–3018, May 2014.

[25] P. Stoica and M. Viberg, "Maximum likelihood parameter and rank estimation in reduced-rank multivariate linear regressions ," *IEEE Trans. Sig. Process.*, vol. 44, no. 12, pp. 3069–3078, Dec. 1996.

[26] K. Konstantinides and K. Yao, "Statistical analysis of effective singular values in matrix rank determination ," *IEEE Trans. Acoust., Speech, Sig. Process.*, vol. 36, no. 5, pp. 757–763, 1988.

[27] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series and products*, New York: Academic Press, 1966.

[28] B. Levy, *Principles of Signal Detection and Parameter Estimation*. New York: Springer, 2008.

[29] A. Sibille, C. Oestges, and A. Zanella, *MIMO: from theory to implementation*, Academic Press, 2010.

[30] J. R. Hampton, *Introduction to MIMO Communications*, Cambridge university press, 2013.

[31] E. C. van der Meulen, "Three-terminal communication channels," *Advanced Applied Prob.*, vol. 3, pp. 120-154, 1971.

[32] T. M. Cover and A. A. El Gamal, "Capacity theorems for the relay channel," *IEEE Trans. Info. Theory*, vol. 25, pp. 572-184, 1979.

[33] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior,"*IEEE Trans. Info. Theory*, vol. 50, pp. 3062–3080, 2004.

[34] Y. Jing and H. Jafarkhani, "Network beamforming using relays with perfect channel information," *IEEE Trans. Info. Theory*, vol. 55, no. 6, pp. 2499–2517, Jun. 2009.

[35] P. Larsson, "Large-scale cooperative relaying network with optimal coherent combining under aggregate relay power constraints," in *Proc. Future Tele. Conf.*, Beijing, China, Dec. 2003.

[36] H. A. Suraweera, H. Q. Ngo, T. Q. Duong, C. Yuen, and E. G. Larsson, "Multi-pair amplify-and-forward relaying with very large antenna arrays," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2013, pp. 4635–4640.

[37] G. Amarasuriya, "Sum rate analysis for multi-user massive MIMO relay networks," in *Proc. 2015 IEEE Global Commun. Conf.*, San Diego, CA, 2015, pp. 1-7.

[38] G. Amarasuriya and H. V. Poor, "Multi-user relay networks with massive MIMO," in *Proc. 2015 IEEE Int. Conf. Commun.*, London, 2015, pp. 2017-2023.

[39] G. Amarasuriya and H. V. Poor, "Impact of channel aging in multi-way relay networks with massive MIMO," in *Proc. 2015 IEEE Int. Conf. Commun.* , London, 2015, pp. 1951-1957.

[40] H. Cui, L. Song, and B. Jiao, "Multi-pair two-way amplify-and-forward relaying with very large number of relay antennas," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2636-2645, May 2014.

[41] S. Jin, X. Liang, K. K. Wong, X. Gao and Q. Zhu, "Ergodic rate analysis for multipair massive MIMO two-way relay networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1480-1491, Mar. 2015.

[42] X. Jia, P. Deng, L. Yang and H. Zhu, "Spectrum and energy efficiencies for multiuser pairs massive MIMO systems with full-duplex amplify-and-forward relay," *IEEE Access*, vol. 3, pp. 1907-1918, 2015.

[43] T. V. T. Le and Y. H. Kim, "Power and spectral efficiency of multi-pair massive antenna relaying systems with zero-forcing relay beamforming," *IEEE Commun. Letters*, vol. 19, no. 2, pp. 243-246, Feb. 2015.

[44] Y. Wang, S. Li, C. Li, Y. Huang and L. Yang, "Ergodic rate analysis for massive MIMO relay systems with multi-pair users under imperfect CSI," in *2015 IEEE Global Conf. on Sig. and Info. Process.*, Orlando, FL, 2015, pp. 33-37.

122

[45] H. Wang, J. Ding, J. Yang, X. Gao and Z. Ding, "Spectral and energy efficiency for multi-pair massive MIMO two-way relaying networks with imperfect CSI," in *Proc. 2015 IEEE 82nd Veh. Tech. Conf.*, Boston, MA, Sep. 2015, pp. 1-6.

[46] L. Sun and M. R. McKay, "Opportunistic relaying for MIMO wireless communication: relay selection and capacity scaling laws," *IEEE Trans. Wireless Commun.*, vol. 10, no. 6, pp. 1786–1797, Apr. 2011.

[47] V. H. Nassab, S. Shahbazpanahi, A. Grami, and Z. Luo, "Distributed beamforming for relay networks based on second-order statistics of the channel state information," *IEEE Trans. Sig. Process.*, vol. 56, no. 9, pp. 4306–4316, Sept. 2008.

[48] Y. Rong, "Robust design for linear non-regenerative mimo relays with imperfect channel state information," *IEEE Trans. Sig. Process.*, vol. 59, no. 5, pp. 2455–2460, May 2011.

[49] S. Ren and M. van der Schaar, "Distributed power allocation in multi-user multi-channel cellular relay networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 6, pp. 1952–1964, Jun. 2010.

[50] Q. Cao, H. V. Zhao, and Y. Jing, "Power allocation and pricing in multiuser relay networks using stackelberg and bargaining games," *IEEE Trans. Veh. Technol.*, vol. 61, no. 7, pp. 3177–3190, Sep. 2012.

[51] Q. Cao, Y. Jing, and H. V. Zhao, "Power bargaining in multi-source relay networks," in *Proc. IEEE Commun. Conf.*, Jun. 2012, pp. 3905–3909.

[52] H. Yu, R. Xiao, Y. Li, and J. Wang, "Energy-efficient multi-user relay networks," in *Proc. IEEE Wireless Commun. and Sig. Process. Conf.*, Nov. 2011, pp. 1–5.

[53] K. T. Phan, T. Le-Ngoc, S. A. Vorobyov, and C. Tellambura, "Power allocation in wireless multi-user relay networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 5, pp. 2535–2545, May 2009.

[54] E. Koyuncu and H. Jafarkhani, "Distributed beamforming in wireless multiuser relay-interference networks with quantized feedback," *IEEE Trans. Info. Theory*, vol. 58, no. 7, pp. 4538–4576, Jul. 2012.

[55] Z. Bai, R. Li, H. Zhang, and K. Kwak, "Relay power allocation schemes for multiuser cooperative communication," in *Proc. IEEE Wireless Commun. and Networking Conf.*, Mar. 2011, pp. 1612–1616.

[56] H. H. Nguyen, D. H. N. Nguyen, "Power allocation in wireless multiuser multi-relay networks with distributed beamforming," *IET Commun.*, vol. 5, no. 14, pp. 2040–2051, Sept. 2011.

[57] Y. Zhang, X. Li, and M. Amin, "Distributed beamforming in multi-user cooperative wireless networks," in *Proc. IEEE Commun. and Networking in China Conf.*, Aug. 2009, pp.1–5.

[58] S. Fazeli-Dehkordy, S. ShahbazPanahi, and S. Gazor, "Multiple peer-to-peer communications using a network of relays," *IEEE Trans. Sig. Process.*, vol. 57, no. 8, pp. 3053–3062, Aug. 2009.

[59] S. Verdú, *Multiuser Detection*, Cambridge University Press, 1998.

[60] P. Viswanath and D. N. C. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality, " *IEEE Trans. Info. Theory*, vol. 49, no. 8, pp. 1912-1921, Aug. 2003.

[61] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Info. Theory*, vol. 52, no. 9, pp. 3936-3964, Sep. 2006.

[62] H. Yang and T. L. Marzetta, "Performance of conjugate and zero-forcing beamforming in large-scale antenna systems," *IEEE J. Sel. Areas in Commun.*, vol. 31, no. 2, pp. 172-179, Feb. 2013.

[63] H. Q. Ngo and E. G. Larsson, "Large-scale multipair two-way relay networks with distributed AF beamforming," *IEEE Commun. Letters*, vol. 17, no. 12, pp. 2288–2291, Dec. 2013.

[64] S. O. Gharan, A. Bayesteh, and A. K. Khandani, "Asymptotic analysis of amplify and forward relaying in a parallel MIMO relay network," *IEEE Trans. Info. Theory*, vol. 57, no. 4, pp. 2070–2082, Apr. 2011.

[65] C.-K. Wen, K.-K. Wong, and C. T. K. Ng, "On the asymptotic properties of amplify-and-forward MIMO relay channels," *IEEE Trans. Commun.*, vol. 59, no. 2, pp. 590–602, Dec. 2011.

[66] H. Bölcskei, R. U. Nabar, O. Oyman, and A. J. Paulraj, "Capacity scaling laws in MIMO relay networks," *IEEE Trans. Commun.*, vol. 5, no. 6, pp. 1433–1444, Jun. 2006.

[67] J. Wagner, B. Rankov, and A. Wittneben, "Large analysis of amplify-and-forward MIMO relay channels with correlated Rayleigh fading," *IEEE Trans. Info. Theory*, vol. 54, no. 12, pp. 5735–5746, Dec. 2008.

[68] K.-J. Lee, J.-S. Kim, G. Caire, and I. Lee, "Asymptotic ergodic capacity analysis for MIMO amplify-and-forward relay networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 9, pp. 2712–2717, Jul. 2010.

[69] N. Fawaz, K. Zarifi, M. Debbah, and D. Gesbert, "Asymptotic capacity and optimal precoding in MIMO multi-hop relay networks," *IEEE Trans. Info. Theory*, vol. 57, no. 4, pp. 2050–2069, Apr. 2011.

[70] S. Kalyani and R. M. Karthik, "The asymptotic distribution of maxima of independent and identically distributed sums of correlated or non-identical gamma random variables and its applications," *IEEE Trans. Commun.*, vol. 60, no. 9, pp. 2747–2758, Jul. 2012.

[71] R. A. Fisher and L. H. C. Tippett, "Limiting forms of the frequency distribution of the largest or smallest member of a sample," *Math. Proc. Cambridge Philosph. Soc.*, vol. 24, no. 2, pp. 180–190, 1928.

[72] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Trans. Info. Theory* , vol. 51, no. 2, pp. 506–522, Feb. 2005.

[73] M. Xia and S. Aïssa, "Spectrum-sharing multi-hop cooperative relaying: Performance analysis using extreme value theory," *arXiv preprint arXiv:1311.5184*, Nov. 2013.

[74] *The Wolfram functions site* [online]. Available: http://functions.wolfram.com

[75] E. W. Weisstein, "Gumbel Distribution," *MathWorld* [online]. http://mathworld.wolfram.com/GumbelDistribution.html

[76] M. O. Hasna and M. S. Alouini, "Harmonic mean and end-to-end performance of transmission systems with relays," *IEEE Trans. Commun.*, vol. 52, no. 1, pp. 130–135, Mar. 2004.

[77] X. Tang and Y. Hua, "Optimal design of non-regenerative mimo wireless relays," *IEEE Trans. Wireless Commun.*, vol. 6, no. 4, pp. 1398–1407, Apr. 2007.

[78] V. H. Nassab, S. Shahbazpanahi, and A. Grami, "Optimal distributed beamforming for two-way relay networks," *IEEE Trans. Sig. Process.*, vol. 58, no. 3, pp. 1238–1250, Mar. 2010.

[79] H. Chen, A. B. Gershman, and S. Shahbazpanahi, "Filter-and-forward distributed beamforming in relay networks with frequency selective fading," *IEEE Trans. Sig. Process.*, vol. 58, no. 3, pp. 1251–1262, Mar. 2010.

[80] A. El-Keyi and B. Champagne, "Adaptive linearly constrained minimum variance beamforming for multiuser cooperative relaying using the kalman filter," *IEEE Trans. Wireless Commun.*, vol. 9, no. 2, pp. 641–651, Feb. 2010.

[81] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge Univ. Press, 2004.

[82] H. Q. Ngo, E. G. Larsson, and T. L. Thomas, "The multi-cell multi-user MIMO uplink with very large antenna arrays and a finite-dimensional channel," *IEEE Trans. Commun.*, vol. 61, no. 6, pp. 2350–2361, Jun. 2013.

[83] G. Zhu, C. Zhong, H. A. Suraweera, Z. Zhang, C. Yuen, and R. Yin, "Ergodic capacity comparison of different relay precoding schemes in dual-hop AF systems with co-channel interference," *IEEE Trans. Wireless Commun.*, vol. 62, no. 7, pp. 2314-2328, Jul. 2014.

[84] G. Zhu, C. Zhong, H. A. Suraweera, Z. Zhang, and C. Yuen, "Outage probability of dual-hop multiple antenna AF systems with linear processing in the presence of co-channel interference," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2308-2321, Apr. 2014.

[85] H. Q. Ngo, E. G. Larsson and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436-1449, Apr. 2013.

[86] Q. Zhang, S. Jin, K. K. Wong, H. Zhu and M. Matthaiou, "Power scaling of uplink massive MIMO systems with arbitrary-rank channel means," *IEEE J. Sel. Topics in Sig. Process.*, vol. 8, no. 5, pp. 966-981, Oct. 2014.

[87] W. Zhang, C. Pan, B. Du, M. Chen, X. Gong, and J. Dai, "Downlink SINR study in multiuser large scale antenna systems", *Wireless Personal Commun.*, vol. 79, no. 2, pp. 1539-1556, Nov. 2014.

[88] C. Feng, Y. Jing and S. Jin, "Interference and outage probability analysis for massive MIMO downlink with MF precoding," *IEEE Sig. Process. Letters*, vol. 23, no. 3, pp. 366-370, Mar. 2016.

[89] M. S. Zia and S. A. Hassan, "Outage analysis of multi-user massive MIMO systems subject to composite fading," in *IEEE Veh. Technol. Conf.*, Glasgow, 2015, pp. 1-5.

[90] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590-3600, Nov. 2010.

[91] H. Kobayashi, B. L. Mark, and W. Turin, *Probability, random processes, and statistical analysis*, Cambridge Univ. Press, 2012.

[92] E. Björnson, E. G. Larsson and T. L. Marzetta, "Massive MIMO: ten myths and one critical question," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 114-123, Feb. 2016.

[93] B. M. Hochwald, T. L. Marzetta and V. Tarokh, "Multiple-antenna channel hardening and its implications for rate feedback and scheduling," *IEEE Trans. Info. Theory*, vol. 50, no. 9, pp. 1893-1909, Sept. 2004.

[94] B. Everitt, *The Cambridge Dictionary of Statistics*, Cambridge University Press, 1998.

[95] M. Abramowitz and I. A. Stegunhor, *Handbook of mathematical functions*, Dover, 1964.

[96] M. S. Alouini, A. Abdi, and M. Kaveh, "Sum of gamma variates and performance of wireless communication systems over Nakagami-fading channels," IEEE Trans. Veh. Technol., vol. 50, pp. 1471-1480, Nov. 2001.

[97] B. Hassibi and B. Hochwald, "How much training is needed in multiple-antenna wireless links?", *IEEE Trans. Info. Theory*, vol. 49, pp. 951–961, Apr. 2003.

[98] N. Shariati, J. Wang and M. Bengtsson, "Robust training sequence design for correlated MIMO channel estimation", *IEEE Trans. Sig. Process.*, vol. 61, no. 1, pp. 107–120, 2014.

[99] C. Huang, T. Chang, X. Zhou and Y. Hong, "Two-way training for discriminatory channel estimation in wireless MIMO systems", *IEEE Trans. Sig. Process.*, vol. 61, no. 10, pp. 2724–2738, May 2013.

[100] M. Biguesh, A. Gershman, "Training-based MIMO channel estimation: A study of estimator tradeoffs and optimal training signals", *IEEE Trans. Sig. Process.*, vol. 54, no. 3, pp. 884–893, Mar. 2006.

[101] Y. Jing and X. Yu, "ML-based channel estimations for non-regenerative relay networks with multiple transmit and receive antennas," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 8, pp. 1428–1439, Sept. 2012.

[102] E. Lindskog and C. Tidestav, "Reduced rank channel estimation ," *Proc. IEEE VTC 1999*, Jul. 1999.

[103] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, 1936.

[104] A. Ordonez, D. Palomar and J. Fonollosa, "Ordered eigenvalues of a general class of Hermitian matrices with application to the performance analysis of MIMO systems," *IEEE Trans. Sig. Process.*, vol. 57, no. 2, pp. 672–689, Feb. 2009.

[105] Y. Hua, M. Nikpour and P. Stoica, "Optimal reduced-rank estimation and filtering," *IEEE Trans. Sig. Process.*, vol. 49, no. 3, pp. 457–469, Mar. 2001.

[106] R. M. Corless, G. H. Gonnet, D. E. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Advances in Comput. Math.*, vol. 5, no. 1, pp. 329–359, Dec. 1996.

[107] D. A. Barry, J. Y. Parlange, L. Li, H. Prommer, C. J. Cunningham, and F. Stagnitti, "Analytical approximations for real values of the Lambert W-function," *Math. and Comput. in Simulation*, vol. 53, no. 1, pp. 95–103, Aug. 2000.

[108] M. O. Hasna and M. S. Alouini, "End-to-end performance of transmission systems with relays over rayleigh-fading channels," *IEEE Trans. Wireless Commun.*, vol. 2, no. 6, pp. 1126–1131, Nov. 2003.

[109] J. S. Goldstein and I. S. Reed and L. L. Scharf, "A multistage representation of the Wiener filter based on orthogonal projections," *IEEE Trans. Info. Theory*, vol. 44, no. 7, pp. 2943-2959, Nov. 1998.

[110] M. L. Honig and J. S. Goldstein, "Adaptive reduced-rank interference suppression based on the multistage Wiener filter," *IEEE Trans. Commun.*, vol. 50, no. 6, pp. 986-994, Jun. 2002.

[111] R. C. de Lamare and M. Haardt and R. Sampaio-Neto, "Blind adaptive constrained reduced-rank parameter estimation based on constant modulus design for CDMA interference suppression," *IEEE Trans. Sig. Process.*, vol. 56, no. 6, pp. 2470-2482, Jun. 2008.

[112] Y. Sun and V. Tripathi and M. L. Honig, "Adaptive turbo reduced-rank equalization for MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2789-2800, Nov. 2005.

[113] H. Qian and S. N. Batalama, "Data record-based criteria for the selection of an auxiliary vector estimator of the MMSE/MVDR filter," *IEEE Trans. Commun.*, vol. 51, no. 10, pp. 1700-1708, Oct. 2003.

[114] R. C. de Lamare and R. Sampaio-Neto, "Adaptive reduced-rank processing based on joint and iterative interpolation, decimation, and filtering," *IEEE Trans. Sig. Process.*, vol. 57, no. 7, pp. 2503-2514, Jul. 2009.

[115] R. C. de Lamare and R. Sampaio-Neto, "Adaptive reduced-rank equalization algorithms based on alternating optimization design techniques for MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 60, no. 6, pp. 2482-2494, Jul. 2011.

[116] M. Nicoli, "Multi-user reduced rank receivers for TD/CDMA systems", Ph.D. dissertation, Politecnico di Milano, Milan, Italy, Dec. 2001.

[117] T. W. Anderson, "An introduction to multivariate statistical analysis," 2nd ed. *New York: Wiley*, 1984.

[118] C. R. Rao, "Linear statistical inference and its applications," *New York: Wiley*, 1973.

[119] A. G. Burr, "Capacity bounds and estimates for the finite scatterers MIMO wireless channel," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 5, pp. 812–818, Jun. 2003.

[120] R. A. Horn and C. R. Johnson, *Matrix analysis*, *Cambridge University Press*, Feb. 1990.

[121] Q. Wang and Y. Jing, "Power allocation and sum-rate analysis for multi-user multi-relay networks," in *Proc. IEEE Veh. Technol. Conf.*, Las Vegas, NV, Sep. 2013, pp. 1-5.

[122] A. Agustin and J. Vidal, "Amplify-and-forward cooperation under interference-limited spatial reuse of the relay slot," *IEEE Trans. Wireless Commun.*, vol. 7, pp. 1952-1962, May 2008.

[123] X. Zhang, A. F. Molisch, and S. Y. Kung, "Variable phase shift based RF baseband codesign for MIMO antenna selection," *IEEE Trans. Sig. Process.*, vol. 53, no. 11, pp. 4091–4103, Nov. 2005.

[124] P. Karamalis, N. Skentos, and A. G. Kanatas, "Adaptive antenna subarray formation for MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 5, no. 11, pp. 2977-2982, Nov. 2006.

[125] V. Venkateswaran, and A. van der Veen, "Analog beamforming in MIMO communications with phase shift networks and online channel estimation," *IEEE Trans. Sig. Process.*, vol. 58, no. 8, pp. 4131-4143, Aug. 2010.

[126] O. El Ayach, R. W. Heath, Jr., S. Abu-Surra, S. Rajagopal, and Z. Pi, "Low complexity precoding for large millimeter wave MIMO systems," in *Proc. IEEE Int. Conf. Commun.*, Ottawa, Canada, June 2012.

[127] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, Jr., "Hybrid precoding for millimeter wave cellular systems with partial channel knowledge," in *Proc. Info. Theory and Appl. Workshop*, Feb. 2013.

[128] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, Jr., "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499-1513, Mar. 2014.

[129] L. Liang, W. Xu, X. Dong, "Low-complexity hybrid precoding in massive multiuser MIMO systems," *IEEE Wireless Commun. Letters*, vol. 3, no. 6, pp. 653–656, Dec. 2014

[130] M. Kim, Y. Lee, "MSE-based hybrid RF/baseband processing for millimeter wave communication systems in MIMO interference channels," *IEEE Trans. Veh. Technol.* , vol.64, no.6, pp. 2714-2720, Jun. 2015.

[131] J. Lee and Y. H. Lee, "AF relaying for millimeter wave communication systems with hybrid RF/baseband MIMO processing," in *Proc. IEEE Int. Conf. Commun.*, Sydney, NSW, Jun. 2014, pp. 5838–5842.

[132] Q. Wang and Y. Jing, "Closed-form average SNR and ergodic capacity approximations for best relay selection," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2827-2833, Apr. 2016.

[133] Q. Wang and Y. Jing, "Multiple-threshold rank-detection for reduced-rank MIMO channels," in *Proc. IEEE Pacific Rim Conf. Commun. Comput. and Sig. Process.*, Victoria, BC, 2015, pp. 125-130.

[134] Q. Wang and Y.Jing, "New rank detection methods for reduced-rank MIMO systems," *EURASIP J. Wireless Commun. and Networking*, Oct. 2015.

$\sim$

Appendix A

Proof of Lemma 1

With the exact close-form $F(x)$ and $f(x)$ expressions in (3.5) and (3.6), we need to verify that

$$\lim_{x \to \infty} \frac{f'(x)[1 - F(x)]}{f^2(x)} = -1.$$

First, we derive $f'(x)$ as

$$f'(x) = 2\sqrt{\eta} \left(-1 - \eta + \eta P \sigma_f^2 - 2(\eta + 1)x\right) K_0 \left(\frac{2\sqrt{x(1+x)}}{\sqrt{\eta} P \sigma_f^2}\right) \frac{2 \exp\left(\frac{-x(1+\eta)}{\eta P \sigma_f^2}\right)}{\eta^{5/2} (P \sigma_f^2)^3}$$

$$-\frac{\eta + x + \eta(6 + \eta)x + (1 + 6\eta + \eta^2)x^2}{\sqrt{x(x+1)}} K_1 \left(\frac{2\sqrt{x(1+x)}}{\sqrt{\eta} P \sigma_f^2}\right) \frac{2 \exp\left(\frac{-x(1+\eta)}{\eta P \sigma_f^2}\right)}{\eta^{5/2} (P \sigma_f^2)^3}.$$

By noticing that $\exp\left(\frac{-x(1+\eta)}{\eta P \sigma_f^2}\right)$ is the common factor of $1 - F(x)$, $f'(x)$ and $f(x)$, it can be eliminated from both the numerator and denominator. From [74, 03.04.06.0010.01], when $x \gg 1$,

$$K_\nu(x) = \sqrt{\frac{\pi}{2}} \frac{\exp(-x)}{\sqrt{x}} \left(1 + \mathcal{O}\left(\frac{1}{x}\right)\right) \quad \text{for } v = 0, 1.$$

Thus we have

$$\lim_{x \to \infty} \frac{K_0\left(\frac{2\sqrt{x(1+x)}}{\sqrt{\eta}\sigma_f^2 P}\right)}{K_1\left(\frac{2\sqrt{x(1+x)}}{\sqrt{\eta}\sigma_f^2 P}\right)} = 1.$$

After cancelling common factors from the numerator and denominator, we have

$$\lim_{x\to\infty}\frac{f'(x)[1-F(x)]}{f^2(x)}$$

$$=\frac{\frac{2}{\eta^{5/2}(P\sigma_f^2)^3}\left[2\sqrt{\eta}\left(-1-\eta+\eta P\sigma_f^2-2(\eta+1)x\right)-\frac{\eta+x+\eta(6+\eta)x+(1+6\eta+\eta^2)x^2}{\sqrt{x(x+1)}}\right]\frac{2\sqrt{x(1+x)}}{\sqrt{\eta}P\sigma_f^2}}{\left(\frac{2}{\eta^{3/2}(P\sigma_f^2)^2}\left((1+\eta)\sqrt{x(1+x)}+\sqrt{\eta}(1+2x)\right)\right)^2}$$

$$=\frac{\frac{2\sqrt{\eta}\left(-1-\eta+\eta P\sigma_f^2-2(\eta+1)x\right)}{\sqrt{x(x+1)}}-\frac{\eta+x+\eta(6+\eta)x+(1+6\eta+\eta^2)x^2}{x(x+1)}}{\left((1+\eta)+\frac{\sqrt{\eta}(1+2x)}{\sqrt{x(1+x)}}\right)^2}$$

$$=-\frac{4\sqrt{\eta}(\eta+1)+\eta^2+6\eta+1}{(1+\eta+2\sqrt{\eta})^2}=-1.$$

Thus the condition in (3.3) is satisfied.

Appendix B

Proof for Proposition 1

From the definition of b_N in (3.4) and the $F(x)$ in (3.5), b_N is the solution of the following equation:

$$\exp\left(-\frac{l_a(x)}{p}\right)\frac{l_b(x)}{P}K_1\left(\frac{l_b(x)}{P}\right) = \frac{1}{N}, \tag{B.1}$$

where we have defined $l_a(x) \triangleq x(1+\eta)/(\eta\sigma_f^2)$ and $l_b(x) \triangleq 2\sqrt{x(1+x)}/(\sqrt{\eta}\sigma_f^2)$. When $N \to \infty$, we have $1/N \to 0$. For (B.1) to be satisfied, we need $l_a(x)/P \to \infty$ or $l_b(x)/P \to \infty$. Since η is bounded and P is large, this leads to $x/P \to \infty$. Thus the solution of (B.1) is large and has a higher order than P, i.e., $x \gg P$ or equivalently $x/P \gg 1$. So when $P \gg 1$, we have $x + 1 \approx x$, and $l_b(x) \approx 2x/(\sqrt{\eta}\sigma_f^2)$. The equation in (B.1) can be approximated as

$$\exp\left(-\frac{l_a(x)}{P}\right)\frac{2x}{\sqrt{\eta}\sigma_f^2}\frac{1}{P}K_1\left(\frac{2x}{\sqrt{\eta}\sigma_f^2}\frac{1}{P}\right) = \frac{1}{N}. \tag{B.2}$$

According to [74, 03.04.06.0010.01], when $x/P \gg 1$,

$$K_1\left(\frac{2x}{\sqrt{\eta}\sigma_f^2}\frac{1}{P}\right) = \sqrt{\frac{\pi}{2}}\frac{\exp\left(-\frac{2x}{\sqrt{\eta}\sigma_f^2}\frac{1}{P}\right)}{\sqrt{\frac{2x}{\sqrt{\eta}\sigma_f^2}\frac{1}{P}}}\left(1 + \mathcal{O}\left(\frac{1}{\frac{2x}{\sqrt{\eta}\sigma_f^2}\frac{1}{P}}\right)\right)$$

$$\approx \frac{\sqrt{\pi}}{2}\eta^{\frac{1}{4}}\sigma_f\sqrt{\frac{P}{x}}\exp\left(-\frac{2}{\sqrt{\eta}\sigma_f^2}\frac{x}{P}\right). \tag{B.3}$$

Applying this approximation to (B.2), after some straightforward manipulations, we get

$$\frac{-2(\sqrt{\eta}+1)^2}{\sigma_f^2\eta}\frac{x}{P}\exp\left(\frac{-2(\sqrt{\eta}+1)^2}{\sigma_f^2\eta}\frac{x}{P}\right) = \frac{-2(\sqrt{\eta}+1)^2}{N^2\pi\sqrt{\eta}}. \tag{B.4}$$

The inverse function for $z \exp(z)$ (when $z < -1/e$) is the Lambert W function $W_{-1}(x)$ [106]. Notice that since $x/P \gg 1$, the condition $\frac{-2(\sqrt{\eta}+1)^2}{\sigma_f^2 \eta} \frac{x}{P} < -\frac{1}{e}$ is satisfied. Thus the solution for Equation (B.4) for large P can be approximated as follows:

$$b_N \approx -\frac{\sigma_f^2 \eta}{2(\sqrt{\eta}+1)^2} W_{-1}\left(\frac{-2(\sqrt{\eta}+1)^2}{\pi\sqrt{\eta}} \frac{1}{N^2}\right) P. \tag{B.5}$$

Since this solution of b_N is still not favourable for analysis, we use the following tight approximation for $W_{-1}(x)$ [107]:

$$W_{-1}(x) \approx \ln(-x) - 5.9506\left[1 - \left(1 + \frac{0.2377\sqrt{-\ln(-ex)}}{1 + 0.0042\ln(-ex)\exp\left(-0.0201\sqrt{-\ln(-ex)}\right)}\right)^{-1}\right]. \tag{B.6}$$

By using (B.6) in (B.5), we have

$$b_N \approx \sigma_f^2 \sqrt{\eta} C_1^2 \left[\ln(NC_1) + \ln\left(\sqrt{\frac{\pi}{2}}\right) + C_2\right] P, \tag{B.7}$$

where C_1 and C_2 are defined in (3.9) and (3.10).

For a_N, from (3.4) and the facts that $b_N \gg P$, $b_N + 1 \approx b_N$, and

$$K_0\left(\frac{2b_N}{\sqrt{\eta}P\sigma_f^2}\right) / K_1\left(\frac{2b_N}{\sqrt{\eta}P\sigma_f^2}\right) \approx 1,$$

we have

$$a_N \approx \sigma_f^2 \sqrt{\eta} C_1^2 P. \tag{B.8}$$

By substituting (B.7) and (B.8) in (3.8), after some straightforward manipulations, the CDF in (3.11) can be obtained. The PDF is derived from the derivative of the CDF.

Appendix C

Proof for Theorem 1

First, a widely-used SNR upper bound [14] is

$$\text{SNR}_m < \text{SNR}_{ub} = \sum_{n=1}^{N} \frac{\gamma_m PQ|f_{mn}g_{nm}|^2}{|f_{mn}|^2 P + \gamma_m|g_{nm}|^2 Q}.$$

Notice that the terms in the summation series are i.i.d.. When $N \to \infty$, from the law of large numbers,

$$\frac{1}{N}\text{SNR}_{ub} \overset{a.s.}{\to} \mathbb{E}\left(I_m\right),$$

where $\mathbb{E}$ stands for the expectation, a.s. stands for almost surely, and

$$I_m \triangleq \frac{\gamma_m PQ|f_{mn}g_{nm}|^2}{|f_{mn}|^2 P + \gamma_m|g_{nm}|^2 Q}.$$

The CDF of I_m is [108]:

$$F_{I_m}(x) = 1 - ax\exp(-bx)K_1(ax),$$

where $a \triangleq 2\eta_m^{-\frac{1}{2}}(\sigma_{f,m}^2 P)^{-1}$, $b \triangleq (\eta_m^{-1} + 1)(\sigma_{f,m}^2 P)^{-1}$, and $K_\nu(\cdot)$ is the ν-th order modified Bessel function of the second kind. From the CDF, the PDF of I_m can be derived as

$$f_{I_m}(x) = abx\exp(-bx)K_1(ax) + ax\exp(-bx)K_0(ax),$$

using which we can show that

$$\mathbb{E}(I_m) = \frac{b(b^2 - a^2) - a^2\sqrt{b^2 - a^2}\ln\left(\frac{b+\sqrt{b^2-a^2}}{a}\right)}{(b^2 - a^2)^2}.$$

Note that $(\eta_m \sigma_{f,m}^2 P)^2 (b^2 - a^2) = (\eta_m - 1)^2$. After straightforward manipulation, we obtain

$$\mathbb{E}(I_m) = P\sigma_{f,m}^2 \frac{\eta_m^3 - 2\eta_m^2 \ln(\eta_m) - \eta_m}{(\eta_m - 1)^3}.$$

Next, we derive a lower bound for SNR. To simplify notation, we define $X \triangleq P|f_{mn}|^2$ and $Y \triangleq \gamma_m Q|g_{nm}|^2$. Then, $X \sim \exp(1/(P\sigma_{f,m}^2))$ and $Y \sim \exp(1/(\gamma_m Q\sigma_{g,m}^2))$. The following inequality always holds

$$(X + Y)^2 > (X + Y)^2 - 1 = (X + Y + 1)(X + Y - 1).$$

Multiplying both sides with $\frac{XY}{(X+Y)^2(X+Y+1)}$, we get

$$\frac{XY}{X + Y + 1} > \frac{XY(X + Y - 1)}{(X + Y)^2} = \frac{XY}{X + Y} - \frac{XY}{(X + Y)^2}$$

Averaging both sides, we have

$$\frac{1}{N}\text{SNR}_m = \mathbb{E}\left[\frac{XY}{X + Y + 1}\right]$$
$$> \mathbb{E}\left[\frac{XY}{X + Y}\right] - \mathbb{E}\left[\frac{XY}{(X + Y)^2}\right] = \mathbb{E}(I_m) - \mathbb{E}[Z],$$

where $Z \triangleq \frac{XY}{(X+Y)^2}$. We can see that $Z \leq 1$. Therefore, $\mathbb{E}(Z) \leq \mathcal{O}(1)$. Then, combining the upper bound and the lower bound, we conclude

$$\mathbb{E}(I_m) - \mathcal{O}(1) < \frac{1}{N}\text{SNR}_m < \mathbb{E}(I_m),$$

which proves (4.8) in Theorem 1.

Appendix D

Proof of Lemma 2

Here, we show the derivations of $\mathbb{E}\{P_{s,e}\}$ and $\mathrm{SCV}\{P_{s,e}\}$. The rest can be derived in a similar way.

Firstly, we have

$$
\mathbb{E}\{P_{s,e}\} = \mathbb{E}\left\{ \frac{|\hat{\mathbf{g}}_i\hat{\mathbf{g}}_i^H\hat{\mathbf{f}}_i^H\hat{\mathbf{f}}_i + \sum_{k=1,k\neq i}^{K} \hat{\mathbf{g}}_i\hat{\mathbf{g}}_k^H\hat{\mathbf{f}}_k^H\hat{\mathbf{f}}_i|^2}{M^4} \right\}
$$

$$
= \mathbb{E}\left\{ \frac{\left(\|\hat{\mathbf{g}}_i\|_F^2\|\hat{\mathbf{f}}_i\|_F^2 + \sum_{k=1,k\neq i}^{K} \hat{\mathbf{g}}_i\hat{\mathbf{g}}_k^H\hat{\mathbf{f}}_k^H\hat{\mathbf{f}}_i\right)\left(\|\hat{\mathbf{g}}_i\|_F^2\|\hat{\mathbf{f}}_i\|_F^2 + \sum_{k=1,k\neq i}^{K} \hat{\mathbf{g}}_i\hat{\mathbf{g}}_k^H\hat{\mathbf{f}}_k^H\hat{\mathbf{f}}_i\right)^H}{M^4} \right\}
$$

$$
= \mathbb{E}\left\{ \frac{\|\hat{\mathbf{g}}_i\|_F^4\|\hat{\mathbf{f}}_i\|_F^4}{M^4} \right\} + \sum_{k=1,k\neq i}^{K} \mathbb{E}\left\{ \frac{|\hat{\mathbf{g}}_i\hat{\mathbf{g}}_k^H\hat{\mathbf{f}}_k^H\hat{\mathbf{f}}_i|^2}{M^4} \right\}, \tag{D.1}
$$

where the last step is obtained because the means of the cross terms are zero.

In the first term of (D.1), as entries of $\hat{\mathbf{g}}_i$ and $\hat{\mathbf{f}}_i$ are i.i.d. whose distribution follows $\mathcal{CN}(0, P_c)$, $\|\hat{\mathbf{g}}_i\|_F^2$ and $\|\hat{\mathbf{f}}_i\|_F^2$ have a gamma distribution with shape parameter M and scale parameter P_c. Thus,

$$
\mathbb{E}\left\{ \frac{\|\hat{\mathbf{g}}_i\|_F^4\|\hat{\mathbf{f}}_i\|_F^4}{M^4} \right\} = P_c^4\left(1 + \frac{2}{M} + \frac{1}{M^2}\right) \approx P_c^4,
$$

where the approximation is made by ignoring lower order terms of M when $M \gg 1$.

In the rest $K - 1$ terms of (D.1), we have

$$
\hat{\mathbf{g}}_i\hat{\mathbf{g}}_k^H\hat{\mathbf{f}}_k^H\hat{\mathbf{f}}_i = \sum_{m_g=1}^{M}\sum_{m_f=1}^{M} \hat{g}_{i,m_g}\hat{g}_{k,m_g}^*\hat{f}_{k,m_f}^*\hat{f}_{i,m_f},
$$

where $\hat{g}_{i,m_g}$ is the (i, m_g)th entry of $\hat{\mathbf{G}}$, and $\hat{f}_{i,m_f}$ is the (i, m_f)th entry of $\hat{\mathbf{F}}$. Thus $\hat{\mathbf{g}}_i \hat{\mathbf{g}}_k^H \hat{\mathbf{f}}_k^H \hat{\mathbf{f}}_i$ can be seen the summation of M^2 terms of i.i.d. random variables, each with mean 0, variance P_c^2. According to CLT, the distribution of $\frac{\hat{\mathbf{g}}_i \hat{\mathbf{g}}_k^H \hat{\mathbf{f}}_k^H \hat{\mathbf{f}}_i}{M}$ converges to $\mathcal{CN}(0, P_c^4)$ when $M \to \infty$. Then $\frac{|\hat{\mathbf{g}}_i \hat{\mathbf{g}}_k^H \hat{\mathbf{f}}_k^H \hat{\mathbf{f}}_i|^2}{M^4}$ has a gamma distribution with shape parameter 1 and scale parameter P_c^4/M^2. Thus, we can obtain

$$\sum_{k=1,k\neq i}^{K} \mathbb{E} \left\{ \frac{|\hat{\mathbf{g}}_i \hat{\mathbf{g}}_k^H \hat{\mathbf{f}}_k^H \hat{\mathbf{f}}_i|^2}{M^4} \right\} = \frac{(K-1)P_c^4}{M^2}.$$

As $M \gg K$, we have $\frac{(K-1)P_c^4}{M^2} \ll P_c^4$. Thus the mean of P_s is P_c^4.

Similarly, we can derive the variance of $P_{s,e}$ as below.

$$\begin{aligned}
\mathrm{Var}\{P_{s,e}\} &= \mathbb{E} \left\{ \frac{|\hat{\mathbf{g}}_i \hat{\mathbf{g}}_i^H \hat{\mathbf{f}}_i^H \hat{\mathbf{f}}_i + \sum_{k=1,k\neq i}^{K} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_k^H \hat{\mathbf{f}}_k^H \hat{\mathbf{f}}_i|^4}{M^8} \right\} - \mathbb{E}\{P_{s,e}\}^2 \\
&\approx \mathbb{E} \left\{ \frac{\|\hat{\mathbf{g}}_i\|_F^8 \|\hat{\mathbf{f}}_i\|_F^8}{M^8} \right\} - P_c^8 \left(1 + \frac{2}{M} + \frac{1}{M^2} \right)^2 \\
&= P_c^8 \frac{(M+3)^2(M+2)^2(M+1)^2}{M^6} - P_c^8 \left(1 + \frac{2}{M} + \frac{1}{M^2} \right)^2 \approx \frac{8P_c^8}{M},
\end{aligned}$$

where the approximations are made by keeping the dominant terms of M. Then, we have $\mathrm{SCV}\{P_{s,e}\} = \mathrm{Var}\{P_{s,e}\}/(\mathbb{E}\{P_{s,e}\})^2 = 8/M$.

Appendix E

Proof of Proposition 4

When $K = 2$, $P_{i,e} = \left| \frac{\mathbf{g}_i \hat{\mathbf{g}}_i^H \hat{\mathbf{f}}_i^H \mathbf{f}_k}{\sqrt{M^3}} + \frac{\mathbf{g}_i \hat{\mathbf{g}}_k^H \hat{\mathbf{f}}_k^H \mathbf{f}_k}{\sqrt{M^3}} \right|^2 / (K - 1)$. Then, using CLT, $P_{i,e}$ has an exponential distribution with parameter $\frac{1}{d_e}$. Then, the PDF can be approximated as $f_{P_{i,e}}(y) \approx \exp(-y/d_e)/d_e$, which is the same as (5.33) for $K = 2$.

Now, we work on the more complicated case of $K \geq 3$. Firstly,

$$\frac{|\mathbf{g}_i \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \mathbf{f}_k|^2}{M^3} = \left| \frac{\mathbf{g}_i \hat{\mathbf{g}}_i^H \hat{\mathbf{f}}_i^H \mathbf{f}_k}{\sqrt{M^3}} + \frac{\mathbf{g}_i \hat{\mathbf{g}}_k^H \hat{\mathbf{f}}_k^H \mathbf{f}_k}{\sqrt{M^3}} + \sum_{n \neq i, n \neq k}^{M} \frac{\mathbf{g}_i \hat{\mathbf{g}}_n^H \hat{\mathbf{f}}_n^H \mathbf{f}_k}{\sqrt{M^3}} \right|^2$$

With the help of CLT, as $M \gg 1$, $\frac{\mathbf{g}_i \hat{\mathbf{g}}_i^H \hat{\mathbf{f}}_i^H \mathbf{f}_k}{\sqrt{M^3}}$ is approximately distributed as $CN(0, P_c^3 + \frac{P_c^2}{M})$, and $\frac{\mathbf{g}_i \hat{\mathbf{g}}_n^H \hat{\mathbf{f}}_n^H \mathbf{f}_k}{\sqrt{M^3}}$ is approximately distributed as $CN(0, \frac{P_c^2}{M})$. We can further show that the covariances between $\frac{\mathbf{g}_i \hat{\mathbf{g}}_i^H \hat{\mathbf{f}}_i^H \mathbf{f}_k}{\sqrt{M^3}}$, $\frac{\mathbf{g}_i \hat{\mathbf{g}}_n^H \hat{\mathbf{f}}_n^H \mathbf{f}_k}{\sqrt{M^3}}$, and $\frac{\mathbf{g}_i \hat{\mathbf{g}}_k^H \hat{\mathbf{f}}_k^H \mathbf{f}_k}{\sqrt{M^3}}$ are zero, thus they are uncorrelated. For tractable analysis, we assume independence as they are Gaussian distributed. Now we conclude that $\frac{|\mathbf{g}_i \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \mathbf{f}_k|^2}{(K-1)M^3}$ has a gamma distribution with shape parameter 1 and scale parameter $\frac{P_c^3}{K-1}\left(2 + \frac{K}{MP_c}\right)$, which is also defined as d_e.

Using CLT, the covariance between $\frac{|\mathbf{g}_i \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \mathbf{f}_k|^2}{(K-1)M^3}$ and $\frac{|\mathbf{g}_i \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \mathbf{f}_l|^2}{(K-1)M^3}$ ($k \neq l$) can be derived as

$$\text{Cov} = \frac{4P_c^5 + 10P_c^6}{(K-1)^2 M} + \frac{18P_c^5 + (2K-4)P_c^6}{(K-1)^2 M^2}, \tag{E.1}$$

where the proof is omitted due to the space limit. The correlation coefficient between the two is subsequently

$$\rho_{jl} = \frac{\text{Cov}\left\{ \frac{|\mathbf{g}_i \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \mathbf{f}_k|^2}{(K-1)M^3}, \frac{|\mathbf{g}_i \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \mathbf{f}_l|^2}{(K-1)M^3} \right\}}{\sqrt{\text{Var}\left\{ \frac{|\mathbf{g}_i \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \mathbf{f}_k|^2}{(K-1)M^3} \right\} \text{Var}\left\{ \frac{|\mathbf{g}_i \hat{\mathbf{G}}^H \hat{\mathbf{F}}^H \mathbf{f}_l|^2}{(K-1)M^3} \right\}}} \approx \frac{1}{M} \frac{\frac{4}{P_c} + 10}{(2 + \frac{K}{MP_c})^2}, \tag{E.2}$$

where the approximation is made by only keeping the dominant terms. It equals ρ_e^2 based on the definition in (5.30).

Thus $P_{i,e}$ is a summation of $K-1$ correlated random variables following the same Gamma distribution. From Corollary 1 of [96], the PDF of $P_{i,e}$ is

$$f_{P_{i,e}}(y) = \prod_{i=1}^{K-1} \left(\frac{\sigma_1}{\sigma_i}\right) \sum_{j=0}^{\infty} \frac{\delta_j y^{K+j-2} \exp\left(-y/\sigma_1\right)}{\sigma_1^{K+j-1}\Gamma(K+j-1)}, \tag{E.3}$$

where $\sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_{K-1}$ are the ordered eigenvalues of the $(K-1)\times(K-1)$ matrix $\mathbf{A}$, whose diagonal entries are d_e and off-diagonal entries are $d_e\rho_e$, and δ_j's are defined iteratively as

$$\delta_0 \triangleq 1, \quad \delta_{j+1} \triangleq \frac{1}{j+1}\sum_{m=1}^{j+1}\left[\sum_{n=1}^{K-1}\left(1-\frac{\sigma_1}{\sigma_n}\right)^m\right]\delta_{j+1-m}. \tag{E.4}$$

As $\mathbf{A}$ is a circulant matrix whose off-diagonal entries are the same, its eigenvalues can be calculated as

$$\sigma_1 = \cdots = \sigma_{K-2} = d_e - d_e\rho_e, \tag{E.5}$$

$$\sigma_{K-1} = d_e + (K-2)d_e\rho_e. \tag{E.6}$$

Then we can show that

$$\delta_j = \left(\frac{(K-1)\rho_e}{1+(K-2)\rho_e}\right)^j = \left(\frac{b_e}{b_e+c_e}\right)^j. \tag{E.7}$$

Substituting (E.5), (E.6) and (E.7) into (E.3), we can get PDF of $P_{i,e}$ as in (5.32) in Proposition 4. Notice that

$$\sum_{i=0}^{\infty}\left(\frac{b_e}{b_e+c_e}\right)^i \phi(y; K+i-1, d_ec_e)$$

$$= \left(\frac{b_e}{b_e+c_e}\right)^{-(K-2)}\frac{\exp\left(-\frac{y}{d_ec_e}\right)}{d_ec_e}\left(\sum_{n=0}^{\infty}\sum_{n=0}^{K-3}\right)\left(\frac{b_e}{d_ec_e(b_e+c_e)}\right)^n\frac{y^n}{n!}.$$

By Taylor series for exponential function and straightforward calculations, we can obtain the closed-form PDF of $P_{i,e}$ in (5.33).

$$\sim$$