# AVERAGE TRANSMIT POWER ANALYSIS AND USER CLUSTERING DESIGN OF DOWNLINK MULTI-ANTENNA NOMA WITH MATCHED FILTER BEAMFORMING

by

Zeyu Sun

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

Master of Science

in

Communications

Department of Electrical and Computer Engineering

University of Alberta

# Abstract

Non-orthogonal multiple access (NOMA) in power domain is considered as a candidate technique in the next generation mobile networks as it can improve the spectral efficiency, the number of connected devices and user fairness compared to traditional orthogonal multiple access (OMA) techniques.

In this thesis we analyze the average transmit power in downlink multi-antenna NOMA systems with two-user clusters where the signal-to-interference-plus-noise-ratios (SINRs) for all users are guaranteed. For systems with a single cluster, a modified NOMA scheme based on the threshold on the alignment of channel directions is proposed to save the transmit power and theoretical analysis in terms of the average transmit power is conducted to demonstrate the superiority of our proposed scheme compared to the original NOMA scheme. To further improve the alignment-based NOMA scheme, a hybrid of NOMA and multi-user beamforming is also proposed. In addition, for systems with more than two users, user clustering algorithms are developed to group the users into multiple two-user clusters with respect to the minimization of the total transmit power. Simulation results validate the correctness of our theoretical results and demonstrate the performance improvement brought by the clustering algorithms.

# Preface

A part of Chapter 2 of the thesis has been published as Z. Sun, Y. Jing and X. Yu, "NOMA Design with Power-Outage Tradeoff for Two-User Systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 8, pp. 1278-1282, Aug. 2020.

Parts of Chapter 2 and Chapter 4 have been published as Z. Sun and Y. Jing, "Average Power Analysis and User Clustering Design for MISO-NOMA Systems" *in Proc. IEEE 21th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, pp. 1-5, 2020.

# Acknowledgements

The completion of this thesis would not have been possible without the following people.

First of all, I would like to express my greatest gratitude to my supervisor Dr. Yindi Jing. It is my luck and honor to work with Dr. Jing. I learned a lot from her, not only on academic studies, but also on how to be a better person.

I would also like to thank my friends: Yuxiang Wang, Danyang Wang, Yitian Zhang and Donghan Li for their help and encouragement during my graduate study. Special thanks to Zijing Niu and Xiang Li for accompanying and encouraging me when I was lost.

Last but not least, I would like to thank my parents Liping Xue and Yonghai Sun for their unconditional supports and understanding.

# Contents

# List of Figures

# Chapter 1

# Introduction

We are living in the information era, where communications through wireless networks becomes an necessity. As an essential infrastructure in modern society, wireless communications provide people with the possibility to interact naturally with anyone from anywhere at any time for anything and has changed people's lives dramatically.

## 1.1 Evolution of wireless communications

The history of wireless communications can be traced back to 1865 when James Clerk Maxwell predicted the existence of electromagnetic wave and proposed the Maxwell's equations in "A Dynamical Theory of the Electromagnetic Field" [1]. In 1888, Heinrich Hertz generated and detected the electromagnetic wave successfully, which demonstrated Maxwell's prediction. The first prototype of wireless communication systems was created by Guglielmo Marconi and he transmitted the first wireless communication message over the Bristol channel in 1897. Later in December 1901, Marconi sent the first oversea message from England and the message was successfully received in Canada.

The first generation (1G) mobile communication appeared in the 1970s when the Bell Lab proposed the concept of cellular network and developed the Advanced Mobile Phone System (AMPS). The AMPS was operated on the 850MHz band with different frequencies for uplink and downlink transmission. There were many variants of AMPS such as the Total Access Communication System (TACS) used Europe and Japan which was operated on the 900MHz

band. Although the first generation mobile communications systems received great success, the drawbacks brought by the analog signal, for example, low spectral efficiency and poor communication security, have to be solved, which motivated the second generation (2G) mobile communication with digital signal.

Typical 2G standards include the Global System for Mobile Communications (GSM) developed by the European Telecommunications Standards Institute (ETSI), the Digital AMPS (D-AMPS) which was a further development of the AMPS system in the US and Canada, and Interim Standard 95 (IS-95) developed by Qualcomm. As the initial 2G networks focused on the voice transmission and data transmission with a low rate, from 1996, standards such as the General Packet Radio Service (GPRS) and IS-95B were proposed to support the data transmission with an improved rate.

The third generation (3G) mobile communications developed for faster data transmission was based on the International Mobile Telecommunications-2000 (IMT-2000) established by the International Telecommunication Union (ITU). The first standard achieved the requirements of IMT-2000 is the Universal Mobile Telecommunications System (UMTS) developed by the 3rd Generation Partnership Project (3GPP) and it was mainly used in Europe, Japan and China. Another 3G standard developed by the 3rd Generation Partnership Project 2 (3GPP2) was Code Division Multiple Access 2000 (CDMA2000), which was a successor of the IS-95 standard and mainly used in North America and South Korea.

The long term evolution (LTE) standard was firstly proposed by NTT Docomo in 2004 and developed by the 3GPP. The LTE Release 8, approved in 3GPP at the end of 2007, was the first release of LTE standard with downlink peak data rate of 300 Mbit/s and uplink peak data rate of 75 Mbit/s. Later in 2008, the ITU Recommendation Sector (ITU-R) established the IMT-Advanced, a successor of IMT-2000, as the requirement for the forth generation (4G) mobile network. Since the LTE Release 8 cannot comply with the requirements in IMT-Advanced such as the peak data rate up to 1 Gbit/s, it was called the 3.9G. The first 4G standard which can meet all the requirements

was LTE Release 10, which was also called LTE-Advanced (LTE-A). With the technologies such as the multiple-input-multiple-output (MIMO) systems, carrier aggression, 256 quadrature amplitude modulation (QAM) and orthogonal frequency division multiplexing (OFDM), LTE-A can reach the downlink peak data rate of 3 Gbit/s and uplink peak data rate of 1.5 Gbit/s.

## 1.2 Next generation mobile communications

The exponentially increasing demand on mobile and wireless communications is propelled by the increasing mobile services and applications. As shown in Fig. 1.1, the overall data traffic is predicted to grow at a compound annual growth rate (CAGR) of 46 percent from 2017 to 2022 and reach 77 exabytes per month by 2022, which achieves an six-fold increase over 2017-2022 [2].

Further, driven by the machine-type applications and the evolution of smartphones, the number of mobile devices also increases rapidly. As shown in Fig. 1.2, the total number of connected devices is excepted to raise from 8.6 billion in 2017 to 12.3 billion in 2022. A significant growth from 11% to 31% on machine-to-machine (M2M) connections can be observed, which also reflect the rapid development of Internet of Things (IoT) industry.

To support the demands of wireless communications, a further evolution on mobile communication system is necessary. The standardization works for the fifth generation (5G) mobile communications are mainly led by ITU, 3GPP and the Institute of Electrical and Electronics Engineers (IEEE). In April 2012, the ITU Recommendation Sector (ITU-R) launched "IMT for 2020 and beyond" program aiming at standardize the requirements for next generation mobile communications, which is referred as IMT-2020. The key capabilities of IMT-2020 are shown in Fig. 1.3, compared with those of IMT-Advanced. Meanwhile, there are also many projects and research activities for the next generation mobile communications across the world, such as the Mobile and wireless communications Enablers for the Twenty-twenty Information Society (METIS), the 5G Public Private Partnership (5G-PPP), the physical layer for dynamic spectrum access and cognitive radio (PHYDYAS), etc..

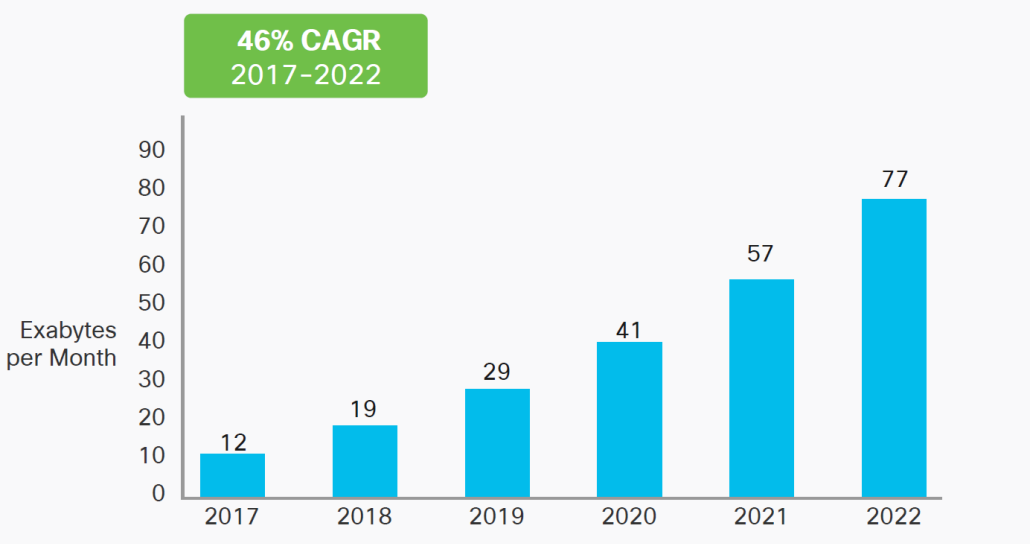


Figure 1.1: Forecasts of mobile data traffic by 2022 (Source: Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022 [2]).

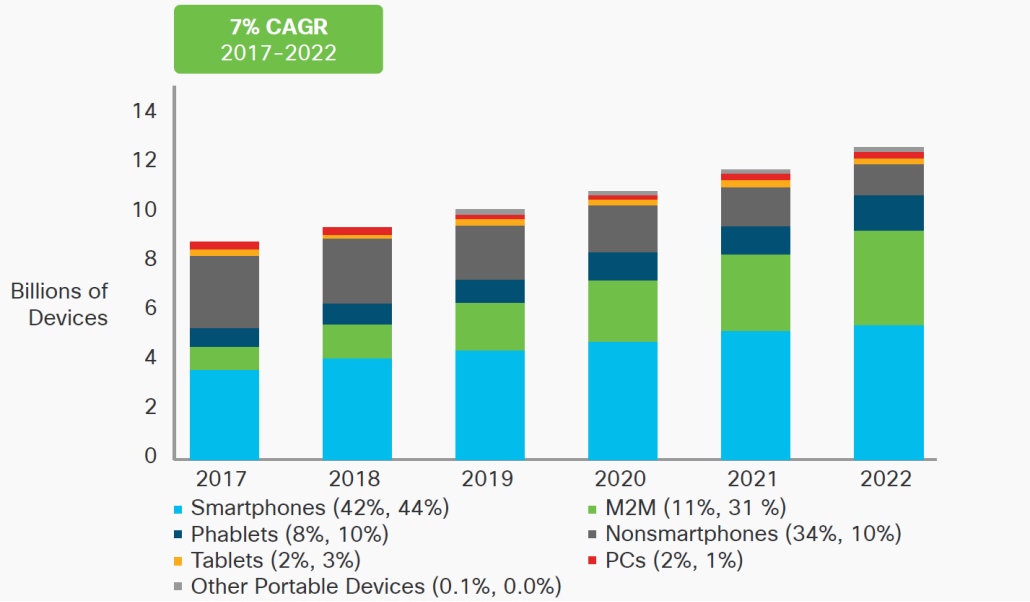


Figure 1.2: Global mobile devices and connections growth (Source: Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022 [2]).

There are many new technologies in 5G networks such as non-orthogonal multiple access (NOMA) [3], mobile edge computing (MEC) [4]–[6], massive MIMO [7]–[10], network slicing [11]–[14], millimeter wave [15]–[18], etc.. Three typical application scenarios defined by ITU-R in 5G network [19] are explained in the following.

- Enhanced Mobile Broadband (eMBB): eMBB provides uniform experience of high data-rate over the coverage area [20]. A typical example of eMBB scenario is the mobile video transmission which is predicted to generate nearly 80% of mobile data traffic by 2022 [2]. Some requirements for eMBB scenario are listed as follows [21]:
  - Peak rate: 20 Gbit/s (downlink) and 10 Gbit/s (uplink);
  - Peak spectral efficiency: 30 bit/s/Hz (downlink) and 15 bit/s/Hz (Uplink);
  - User experience data rate: 100 Mbit/s (downlink) and 50 Mbit/s (uplink);
  - User plane latency: 4 ms.
- Ultra Reliable Low Latency Communications (URLLC): URLLC provides communication services with extremely high reliability and low latency [20]. Some requirements for URLLC scenario are listed as follows [21]:
  - User plane latency: 1 ms;
  - Control plane latency: 10 ms;
  - Reliability: $< 0.001\%$ error probability on the transmission of a layer-2 32-byte protocol data unit within 1 ms.
- Massive Machine Type Communications (mMTC): mMTC provides efficient wireless connectivity for massive low-cost devices and a typical example is the industrial manufacturing [20]. The requirement for mMTC scenario is mainly on the connection density, which is $10^6$ devices/km$^2$ [21].

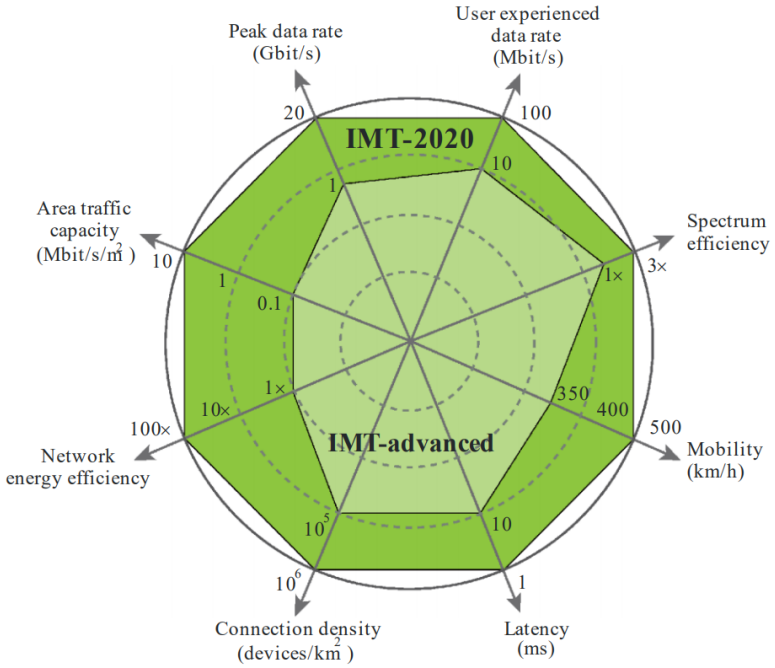


Figure 1.3: Spider diagram for IMT-2020 and IMT-Advanced requirements (Source: ITU-R Recommendation M.2083 [19]).

In practice, these application scenarios do not always appear exclusively. For example, in the communications between autonomous vehicles, the requirement on reliability and latency should be strictly satisfied, which is an URLLC scenario. However, if the vehicles need to exchange video information, a high data rate is also required, which is the eMBB scenario. Therefore, the 5G networks will be flexible with different techniques for different scenarios.

## 1.3 Wireless channel

Different from the wired channel which is stable and predictable, wireless channel is much more complex due to the diverse topography and transmission environment. The fading in wireless environment can be divided into path-loss, shadow fading and multi-path fading according to the causes of fading.

- Path-loss is the degradation of signal power during the propagation of an electromagnetic wave, which increases with the transmitting distance. There are many empirical models for path-loss such as the Hata's model [22] and Lee's model [23].

- When the signal is blocked by obstacles such as hill or large building, there is the shadow of electromagnetic field behind the obstacles, which results in shadow fading. Empirical studies have shown that shadowing fading can be modeled by log-normal distribution [24].

- In wireless environments, the signal can reach the receiver via multiple paths due to the reflection, refraction, diffraction and scattering of the electromagnetic waves. Since the received signal is the summation of signals from different paths and these signals can have different delay, phase and frequency, there is rapid variations in the envelope of the received signal, which is called multi-path fading.

Path-loss and shadow fading reflect the influence of a wireless channel to the signal on large space scale and determine the coverage of the base station (BS) thus called large-scale fading. The multi-path fading can result in rapid fluctuation of signal strength in short distance and time period and thus called small-scale fading.

The small-scale fading can be further categorized into flat-fading and selective-fading. When the signal bandwidth is smaller than the coherent bandwidth which is the inverse of the time spread of multi-path delays, the channel exhibit frequency flat fading; in this case, the transmitted signal in all frequencies experience the same fading. When the signal bandwidth is larger than the coherent bandwidth, the transmitted signal in different frequencies experience different fading, and the channel exhibits frequency selective fading. In this thesis only the flat-fading is considered.

Rayleigh fading and Ricean fading are two common flat fading model. In a rich scattering environment without line-of-sight (LOS) propagation, at any time, the inphase and quadrature component of the received signal are approximately independent and identical distributed (i.i.d.) zero-mean Gaussian random variables with variance $\sigma^2$. As a result, the phase of the received signal at any time is uniformly distributed between $-\pi$ and $\pi$ and its envelope follows the Rayleigh distribution with the following probability density

function (PDF)

$$f_a(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right), \ x \geq 0. \tag{1.1}$$

The average power of the envelope $a$ is given by

$$\mathbb{E}[a^2] = \int_0^\infty x^2 f_a(x) \, \mathrm{d}x = 2\sigma^2.$$

When there exist a LOS path in the environment, at any time $t_1$, the inphase and quadrature components of the received signal follow Gaussian distributions with non-zero means $m_I(t_1)$ and $m_Q(t_1)$ and the same variance $\sigma^2$. In this case, the envelope of the received signal follows Ricean distribution with the following PDF

$$f_a(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2 + A_L^2}{2\sigma^2}\right) I_0\left(\frac{xA_L}{\sigma^2}\right), \ x \geq 0, \tag{1.2}$$

where $I_0(\cdot)$ is the zero-order modified Bessel function of the first kind and $A_L$ is the peak amplitude of the LOS signal.

The Ricean factor $K_R$ is defined as the power ratio of the line-of-sight signal and scattered signal

$$K_R \triangleq \frac{A_L^2}{2\sigma^2}.$$

Obviously, when $K_R = 0$, Ricean distribution becomes Rayleigh distribution; when $K_R \gg 1$, Ricean distribution approaches to Gaussian distribution.

## 1.4 Multiple access techniques in the next generation mobile networks

In a radio cell, multiple users can interfere with each other while accessing to a common BS at the same time using the same frequency. For example, in the downlink, a user receives not only its desired signal from the BS, but also undesired signals sent from the BS to other users. As a result, the performance of the user can be degraded dramatically by the undesired signals. Therefore, to support quality communications of multiple users to access to a common BS, multiple access is proposed.

With traditional multiple access techniques, users access to the BS by orthogonal resource blocks thus called orthogonal multiple access (OMA) scheme.

For example, in frequency division multiple access (FDMA) system, the total bandwidth is divided into non-overlapping frequency bands and each user occupies one of the bands to transmit signals; in time division multiple access (TDMA) system, users transmit signals with different time slots; orthogonal frequency division multiple access (OFDMA) can be regarded as a special case of FDMA where the frequency bands can overlap.

Although the OMA schemes can avoid the interference, the spectral efficiency and the number of served users are limited as each user is only allocated a part of the total resource. Also, when the global network performance is concerned, user fairness is poor because the user with better channel condition is likely to be assigned with more resource than the user with worse channel condition. Therefore, the requirements of IMT-2020, such as the number of connected devices and data rate, may not be achieved by traditional OMA schemes and new multiple access techniques are desirable.

NOMA is a potential technique to help achieve the requirements expected in 5G networks. With NOMA, multiple users can access to a common BS with the same time-frequency resource block without interference by multiplexing users in power domain or code domain, thus the number of connected users and system throughtput can be improved. There are many NOMA schemes such as the power domain NOMA [3], sparse code multiple access (SCMA) [25]–[27], pattern division multiple access (PDMA) [28], [29], interleave division multiple access (IDMA) [30]–[32] and etc. In this thesis, the power domain non-orthogonal multiple access is studied, which is also referred to as NOMA in the following discussion.

## 1.5 Literature review on NOMA systems

The concept of NOMA was firstly proposed in [3] to accommodate the demands on data traffic in future radio access and has attracted considerable research interests in recent years. Early works on NOMA considered the single-input-single-output (SISO) case, e.g., [33]–[35]. For NOMA with multiple uniformly distributed users, in [34], expressions were derived for both the sum-rate and

the outage probability. It was shown that NOMA has higher sum-rate than OMA; while for the outage probability, the choices of user rates and power coefficients are critical. In [33], the sum-rate superiority of NOMA to OMA was shown for the two-user cluster case and fixed power allocation. The significance of user-pairing based on the channel norm difference was also demonstrated. Further, the outage probability of the stronger user was analyzed given signal-to-interference-plus-noise-ratio (SINR) guarantee of the weak user. In [35], for multi-user systems, the sum-rate optimization over power allocation with fairness consideration were studied for both OMA and NOMA systems and the optimized sum-rate of NOMA was shown to be higher.

There are also results on multi-antenna NOMA, where the BS is equipped with multiple antennas, e.g., [36]–[38]. It was proved in [36] that for the two-user case, an upper bound on the sum-rate of OMA also serves as a lower bound on the sum-rate of NOMA when applying the same precoding and postcoding. The work was generalized to the multi-user case in [37]. NOMA systems with a massive antenna array at the BS and multiple 2-user clusters was investigated in [38], where it was showed that multi-user beamforming gives higher average sum-rate than NOMA, but NOMA can outperform when the correlation between the user channels is high, which inspires a hybrid scheme of NOMA and multiuser beamforming.

Obviously, the degree-of-freedom in power domain can be utilized to a greater extent by clustering more users into a single-cluster NOMA system, however, it will increase the complexity which is a main issue in NOMA design [39] and the authors of [37] have proved that adding users into one cluster is detrimental to the sum rate when the total transmit power is fixed.

On the contrary, multi-cluster NOMA system can reduce the complexity of successive interference cancellation (SIC), and is especially beneficial for multi-antenna systems to exploit the degree-of-freedom in the spatial domain [40]. In [41], the authors developed a user pairing algorithm in multi-antenna NOMA system to form multiple two-user clusters, and designed an interference cancellation combining matrix to eliminate the inter-cluster interference. Channel estimation of multi-cluster multi-antenna NOMA system was studied

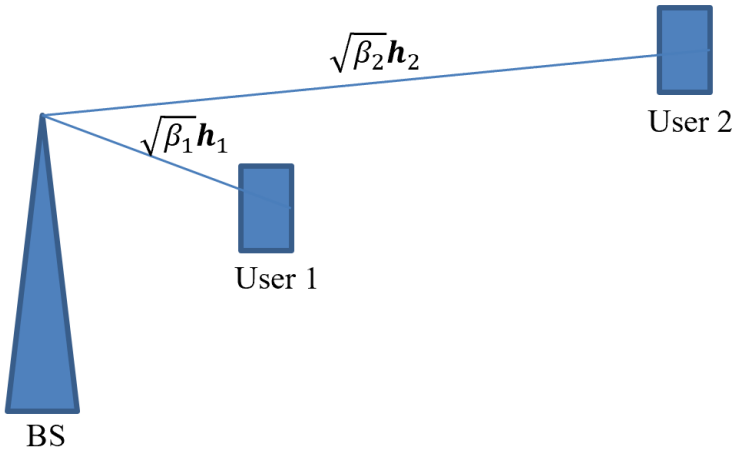


Figure 1.4: The model of a two-user NOMA system.

in [42] where the beamformer for each cluster was designed as a linear combination of user channel vectors. It was concluded that NOMA works well when high quality channel statement information (CSI) is available and the path-loss difference between users in the same cluster is large enough; otherwise multiuser beamforming is preferable.

## 1.6 System model and the principle of NOMA

We consider the downlink transmission from an $M$-antenna BS to two single-antenna users as shown in Fig. 1.4. The channel vector from the BS to User $k$, $\boldsymbol{g}_k \in \mathbb{C}^{M\times 1}$, is modeled as:

$$\boldsymbol{g}_k = \sqrt{\beta_k}\boldsymbol{h}_k, \tag{1.3}$$

where $\beta_k$ is the large-scale fading coefficient; $\boldsymbol{h}_k \in \mathbb{C}^{M\times 1}$ is the i.i.d. Rayleigh fading vector with elements following the circularly symmetric complex Gaussian distribution with zero-mean and unit-variance, i.e.,

$$\boldsymbol{h}_k \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_M), \tag{1.4}$$

where $\boldsymbol{I}_M$ is the identity matrix of size $M$.

In our NOMA system, the BS serves two users with common time-frequency resource block as well as common beamformer. At the BS, the data symbols

for the two users are superposition coded as:

$$s_n = \sqrt{P_{1,n}} s_1 + \sqrt{P_{2,n}} s_2, \tag{1.5}$$

where $s_1, s_2 \sim \mathcal{CN}(0,1)$ are data symbols for the users; $P_{1,n}, P_{2,n}$ are the power allocated to $s_1, s_2$ and the total transmit power is given by $P_n = P_{1,n} + P_{2,n}$; $s_n$ is the transmitted symbol from the BS. Here the subscript $n$ refers to the NOMA scheme.

Let $\boldsymbol{b}_n$ be the BS beamformer. The transmit vector from the BS is thus:

$$\boldsymbol{x}_n = \sqrt{P_{1,n}} \boldsymbol{b}_n s_1 + \sqrt{P_{2,n}} \boldsymbol{b}_n s_2, \tag{1.6}$$

and the signals received by the users are given by:

$$\begin{aligned} y_{1,n} =& \sqrt{P_{1,n}} \boldsymbol{g}_1^H \boldsymbol{b}_n s_1 + \sqrt{P_{2,n}} \boldsymbol{g}_1^H \boldsymbol{b}_n s_2 + n_1, \\ y_{2,n} =& \sqrt{P_{1,n}} \boldsymbol{g}_2^H \boldsymbol{b}_n s_1 + \sqrt{P_{2,n}} \boldsymbol{g}_2^H \boldsymbol{b}_n s_2 + n_2, \end{aligned} \tag{1.7}$$

where $n_k \sim \mathcal{CN}(0,1)$ is the received noise and $(\cdot)^H$ represents the Hermitian of matrix.

According to the principle of SIC, after receiving the signals, one of the users (denote as User $T$, the tail user) decodes its own data symbol $s_T$ by treating the interference as noise; while another user (denote as User $H$, the head user) decodes $s_T$ by treating $s_H$ as noise, then it cancels the component of $s_T$ from $y_{H,n}$ and decodes $s_H$ without interference. The SINR for User $A$ to decode $s_l$ are thus given by:

$$\mathrm{SINR}_{H,s_H} = P_{H,n} \beta_H \left| \boldsymbol{b}_n^H \boldsymbol{h}_H \right|^2, \tag{1.8}$$

$$\mathrm{SINR}_{A,s_T} = \frac{P_{T,n} \beta_A \left| \boldsymbol{b}_n^H \boldsymbol{h}_A \right|^2}{1 + P_{H,n} \beta_A \left| \boldsymbol{b}_n^H \boldsymbol{h}_A \right|^2}, \ A \in \{H, T\}, \tag{1.9}$$

where $|\cdot|$ represents the modulus of complex number.

It can be seen that the two data symbols have different decoding orders, i.e., $s_T$ is always the first symbol to be decoded and $s_H$ is always decoded after $s_T$. Determining the decoding order is equivalent to determining the head and tail users from the two users. Since perfect CSI of both users are assumed at the BS, the BS can decide on the decoding order according to some criteria,

then informs the order to users via a control channel. Here we consider two criteria to determine the decoding order which are referred as Criterion-I and Criterion-II.

With Criterion-I, the head user and tail user are chosen by:

$$\beta_{H_I} \geq \beta_{T_I}, \tag{1.10}$$

where $H_I$ and $T_I$ are the indices of head user and tail user under Criterion-I. In other words, the decoding order with Criterion-I depends on the large-scale fading coefficients exclusively, equivalently, depends on the *average* channel gains. This is different from Criterion-II or the NOMA schemes in [36] and [37], where the decoding order depends on the *instantaneous* channel gains, i.e., both the large-scale fading $\beta_k$'s and the small-scale fading coefficients $\boldsymbol{h}_k$'s. Since the small-scale fading vectors, $\boldsymbol{h}_k$'s, change more frequently than the large-scale fading coefficients, $\beta_k$'s, Criterion-I requires less frequent communications on the decoding order. It is also more robust to the communication error in the control channel.

With Criterion-II, the head user and tail user are chosen by:

$$\beta_{H_{II}} \|\boldsymbol{h}_{H_{II}}\|^2 \geq \beta_{T_{II}} \|\boldsymbol{h}_{T_{II}}\|^2, \tag{1.11}$$

where $\|\cdot\|$ represents the 2-norm of vector. As mentioned, the decoding order with Criterion-II depends on both the large-scale fading $\beta_k$'s and the small-scale fading coefficients $\boldsymbol{h}_k$'s. Although Criterion-II requires more frequent communications via the control channel, it may bring more performance gain since a larger $\beta$ coefficient does not mean larger equivalent channel gain. Criterion-II can guarantee that the decoding order and power allocation are designed based on the equivalent channel gains, which is beneficial to the system performance.

In this thesis, we study the power consumption with the two criteria respectively and the comparison between them is not considered. Such a comparison for uplink NOMA was studied in [43] and the authors showed that Criterion-II is better in terms of outage performance.

## 1.7 Thesis contribution and outline

NOMA is one of the promising technologies for 5G and has gained a lot of research interests in recent years. The achievable sum rate has been one of the most common performance metrics of NOMA systems. However, in most works aiming at comparing or maximizing the sum rate, the QoS of the weaker user cannot be guaranteed since the optimized power allocation schemes assign as much power as possible to the stronger user to maximize the sum rate. Different from these works on the sum-rate and outage probability of NOMA systems, we focus on the power consumption and analyze the required transmit power under SINR constraints for both users.

The thesis contributions are explained in more details by chapters below.

- In Chapter 2, we propose a modified NOMA scheme where NOMA transmissions are conducted only when the the alignment of the channel directions exceeds a threshold and the BS uses matched-filter (MF) precoding along the head user. We derive the instantaneous and average required transmit power to guarantee the SINR levels of both users. Our results show that the average power grows logarithmically in the reciprocal of the alignment threshold and a non-zero threshold is necessary for finite average transmit power. Further, for the scenario that the BS is equipped with massive antenna array, we derive scaling laws of the average transmit power and outage probability with respect to the antenna numbers, as well as their tradeoff law.

- In Chapter 3, we explore the potential of the hybrid of NOMA and multi-user beamforming. We derive the instantaneous required transmit power for multi-user beamforming scheme to guarantee the SINR levels for all users with the same system model in Chapter 1, which shows that a threshold on the alignment of channel directions is necessary when using multi-user beamforming. Then, similiar to Chapter 2, a modified alignment-based multi-user beamforming scheme is proposed and the average required transmit power for this modified scheme to guarantee

the SINR levels for both users is derived. Finally, a hybrid of NOMA and multi-user beamforming based on the alignment threshold is studied to minimize the power consumption and several threshold designs are proposed for the hybrid scheme.

- In Chapter 4, we study the power consumption of multi-cluster multi-antenna NOMA systems. To save transmit power, we design two user-clustering algorithms. One algorithm that achieves the optimal power-minimization finds the solution by solving a matching problem; while the other with suboptimal performance is based on the Hungarian algorithm. Furthermore, we prove analytically that both designs can guarantee the SINR levels for all users with finite average required transmit power when there are more than three users and more than one antenna at the BS.

# Chapter 2

# Average power analysis of the alignment-based NOMA scheme

In this chapter we analyze the average required transmit power of multi-antenna NOMA systems under SINR constraints for both users. First, we derive the instantaneous required transmit power of NOMA scheme to guarantee the SINR levels for both users. Based on the formula of the instantaneous required transmit power and observations from literature, we propose a modified alignment-based NOMA scheme, where NOMA is used only when the alignment of channel directions is above a threshold. The scheme extends the original NOMA and reduces to the original NOMA when the threshold is 0. Then, for the alignment-based NOMA with Criterion-I, a tight approximation is derived for the average required transmit power, which reveals the effect of the alignment threshold and shows the superiority of the alignment-based NOMA to the original NOMA in power saving; further, for the scenario where the BS is equipped with massive antenna array, the scaling laws for the transmit power consumption and the tradeoff between the transmit power and the outage probability are derived, which can be used to guide the threshold design. For the alignment-based NOMA with Criterion-II, the exact expression of average required transmit power is derived and several asymptotic scenarios are discussed. To the best of our knowledge, this work is the first on the average required transmit power analysis for NOMA systems.

The remainder of this chapter is organized as follow. In Section 2.1 we derive the instantaneous required transmit power for NOMA scheme to guar-

antee the SINR levels for both users. Section 2.2 introduce the motivation and mechanism of the alignment-based NOMA scheme; and the average required transmit power for this NOMA scheme with Criterion-I and Criterion-II are investigated in Section 2.3 and Section 2.4, respectively; in Section 2.5 the simulation results are given to validate the theoretical result we derived. Finally, the conclusion of this chapter is given in Section 2.6.

## 2.1 Instantaneous required transmit power with SINR guarantee for both users

We consider a multi-antenna NOMA system with M antennas at the BS and two single-antenna users. The system model and notation have been explained in Section 1.6. Denote the SINR requirements for User $H$ and User $T$ as $\gamma_H$ and $\gamma_T$, respectively. Since $s_T$ needs to be decoded successfully by both users while $s_H$ only needs to be decoded by User $H$, the SINR requirements are formulated as

$$\min\left(\mathrm{SINR}_{H,s_T}, \mathrm{SINR}_{T,s_T}\right) \geq \gamma_T, \tag{2.1}$$

$$\text{and} \quad \mathrm{SINR}_{H,s_H} \geq \gamma_H. \tag{2.2}$$

Recall that $P_{T,n}$ and $P_{H,n}$ are the power coefficients allocated to $s_T$ and $s_H$. By substituting (1.9) into (2.1), we can obtain:

$$\frac{P_{T,n}\min\left(\beta_H\left|\boldsymbol{b}_n^H\boldsymbol{h}_H\right|^2, \beta_T\left|\boldsymbol{b}_n^H\boldsymbol{h}_T\right|^2\right)}{1+P_{H,n}\min\left(\beta_H\left|\boldsymbol{b}_n^H\boldsymbol{h}_H\right|^2, \beta_T\left|\boldsymbol{b}_n^H\boldsymbol{h}_T\right|^2\right)} \geq \gamma_T. \tag{2.3}$$

And a condition of $P_{T,n}$ can be obtained from (2.3), which is given by:

$$P_{T,n} \geq \frac{\gamma_T}{1+\gamma_T}\left(\frac{1}{\min\left(\beta_H\left|\boldsymbol{b}_n^H\boldsymbol{h}_H\right|^2, \beta_T\left|\boldsymbol{b}_n^H\boldsymbol{h}_T\right|^2\right)} + P_n\right). \tag{2.4}$$

By setting the power of the tail user $P_T$ as the lower bound in (2.4), then the SINR for User $H$ to decode $S_H$ can be written as:

$$\begin{aligned}\mathrm{SINR}_{H,s_H} =& P_{H,n}\beta_H\left|\boldsymbol{b}_n^H\boldsymbol{h}_H\right|^2 \\ =& \left[\frac{1}{1+\gamma_T}P_n\beta_H\left|\boldsymbol{b}_n^H\boldsymbol{h}_H\right|^2 - \frac{\gamma_T}{1+\gamma_T}\frac{\beta_H\left|\boldsymbol{b}_n^H\boldsymbol{h}_H\right|^2}{\min\left(\beta_H\left|\boldsymbol{b}_n^H\boldsymbol{h}_H\right|^2, \beta_T\left|\boldsymbol{b}_n^H\boldsymbol{h}_T\right|^2\right)}\right].\end{aligned} \tag{2.5}$$

Finally,

$$P_n \geq P_n^R \triangleq \frac{\gamma_H(1+\gamma_T)}{\beta_H \left|\boldsymbol{b}_n^H \boldsymbol{h}_H\right|^2} + \frac{\gamma_T}{\min\left(\beta_H \left|\boldsymbol{b}_n^H \boldsymbol{h}_H\right|^2, \beta_T \left|\boldsymbol{b}_n^H \boldsymbol{h}_T\right|^2\right)}, \tag{2.6}$$

with equality if and only if (2.1) and (2.2) take equality.

Here in (2.6), $P_n^R$ is the achievable lower bound on the transmit power to guarantee the SINR levels of the two users, which depends on the CSI and the SINR constraints. Also notice that the condition in (2.6) is only a necessary condition since (2.1) and (2.2) may not be satisfied with an inappropriate power allocation scheme, for example, the power allocation schemes in [35] which allocate all the power to the head user in order to maximize the sum rate. On the other hand, if an appropriate power allocation scheme is assumed, the condition in (2.6) is both necessary and sufficient, that is, $\gamma_H$ and $\gamma_T$ are guaranteed for the two users if and only if $P_n \geq P_n^R$.

In this thesis, we adopt the matched filter (MF) beamforming with respect to the head user. The beamformer vector is given by

$$\boldsymbol{b}_n = \frac{\boldsymbol{h}_H}{\|\boldsymbol{h}_H\|}, \tag{2.7}$$

where $\|\cdot\|$ represents the 2-norm of vectors. This is a widely used NOMA beamforming design which has simple implementation and high performance especially for the low SNR region [42]. Moreover, this beamforming scheme can be regarded as a special case of the beamforming scheme proposed in [44] that can create a significant difference between users' channel conditions, which is beneficial to NOMA systems as shown in [33]. The average power consumption, power scaling law, and power-outage tradeoff law which will be studied later in this chapter have not been investigated for this design in the literature.

With the beamformer given by (2.7), (2.6) can be reduced to:

$$P_n^R = \frac{\gamma_H(1+\gamma_T)}{\beta_H \|\boldsymbol{h}_H\|^2} + \frac{\gamma_T}{\min\left(\beta_H \|\boldsymbol{h}_H\|^2, \beta_T \|\boldsymbol{h}_T\|^2 \rho^2\right)}, \tag{2.8}$$

where $\rho$ defined by:

$$\rho \triangleq \frac{\left|\boldsymbol{h}_H^H \boldsymbol{h}_T\right|}{\|\boldsymbol{h}_H\| \|\boldsymbol{h}_T\|} \tag{2.9}$$

is the absolute value of the inner product of the normalized channel vectors of User $H$ and User $T$. In other words, $\rho = |\cos\theta|$, where $\theta$ is the angle between the two channel vectors. To help the presentation, the parameter $\rho$ is referred to as the *alignment of the channel directions* which can measure the similarity between two vectors. For example, when the alignment is 0, the two vectors are orthogonal to each other; and when the alignment is 1, the two vectors are parallel to each other. In some works, $\rho$ is also referred to as channel correlation [38] or absolute correlation coefficient (ACC) in the published version of this chapter.

## 2.2 The alignment-based NOMA scheme and its motivation

As presented in Section 1.6, the two users share a common beam in our multi-antenna NOMA system. However, a single beam may not be able to serve multiple users simultaneously especially when the beam is towards the head user. For example, in Fig 2.1 the channel vectors of the two users are close-to-orthogonal thus the beam towards User 1 cannot serve User 2 while in Fig 2.2 the channels are highly correlated thus the beam can serve both users. The two figures can also be explained by the alignment of channel directions defined in (2.9), i.e., Fig 2.1 is a low alignment scenario while Fig 2.2 is a high alignment scenario.

The significance of the alignment of channel directions can also be observed in (2.8) since the instantaneous required transmit power $P_n^R$ decreases with $\rho$ and grows without limitation when $\rho$ approaches 0, meaning that a high value of $\rho$ is beneficial to power saving. The result in [38] shows that NOMA can outperform multi-user beamforming when $\rho$ is high enough, which demonstrate the importance of $\rho$ again.

Therefore, motivated by the significance of the alignment of channel directions shown above and to save the transmit power of NOMA systems, we propose a modified alignment-based NOMA scheme where the downlink NOMA transmission is only conducted when the the alignment of channel di-

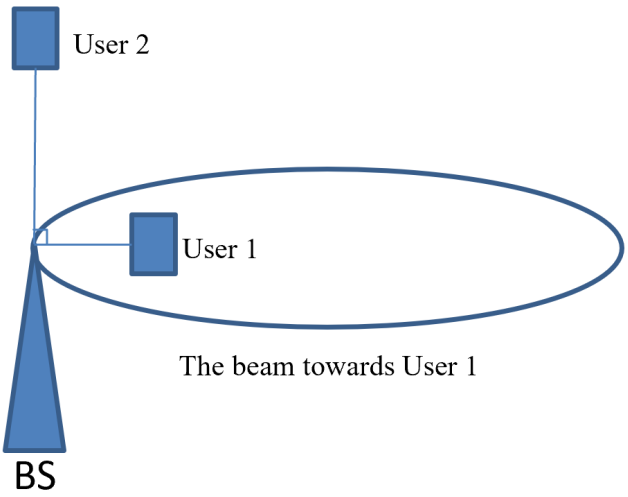


Figure 2.1: Two-user multi-antenna NOMA system with low alignment of channel directions.

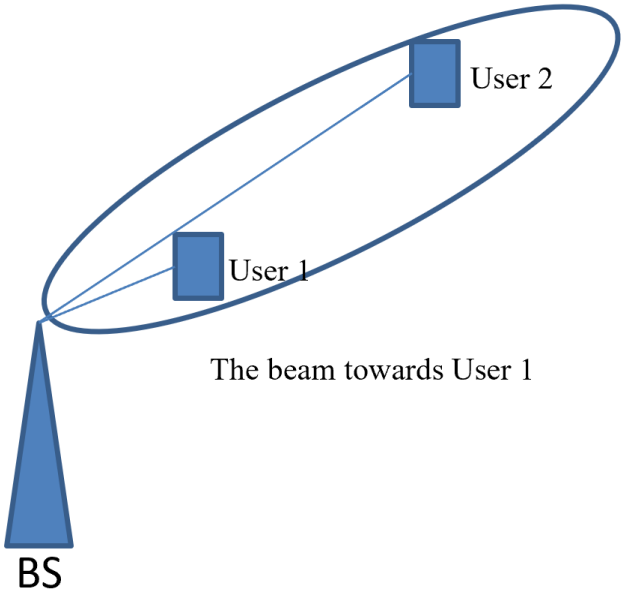


Figure 2.2: Two-user multi-antenna NOMA system with high alignment of channel directions.

rections is above a pre-defined threshold $\rho_{th}$. Otherwise the BS keeps silent.[1] It is straightforward that this modified scheme can save power by reducing the transmissions. On the other hand, it also causes outage for both users when the BS is silent. This modified scheme is a generalization of the original NOMA and reduces to the original NOMA scheme when the alignment-threshold is set to be 0. The design of $\rho_{th}$ and its effects on the performance will be analyzed in the following sections.

## 2.3 Average required transmit power of the alignment-based NOMA with Criterion-I

As explained in Section 1.6 and Section 2.1, the channel vectors are modeled as $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ and $\rho$ is the alignment of channel directions. To help analyze the average required transmit power, we firstly introduce some useful results on the distribution of $\|\boldsymbol{h}_1\|^2$, $\|\boldsymbol{h}_2\|^2$ and $\rho^2$.

**Lemma 2.1** *$\|\boldsymbol{h}_1\|^2$, $\|\boldsymbol{h}_2\|^2$ and $\rho^2$ are mutually independent.*

**Lemma 2.2** *$\|\boldsymbol{h}_1\|^2$ and $\|\boldsymbol{h}_2\|^2$ are independent and identically distributed (i.i.d.) whose PDF is given by:*

$$f_{\|\boldsymbol{h}_1\|^2}(x) = f_{\|\boldsymbol{h}_1\|^2}(x) = \frac{1}{\Gamma(M)} x^{M-1} e^{-x}, \quad x > 0,$$

*which is a Gamma distribution with shape parameter $M$ and scale parameter 1 and $\Gamma(\cdot)$ is the Gamma function.*

**Lemma 2.3** *The PDF of $\rho^2$ is given by:*

$$f_{\rho^2}(x) = (M-1)\,(1-x)^{M-2}, \quad 0 < x < 1,$$

*which is a Beta distribution with shape parameters 1 and $M-1$.*

Lemma 2.1 can be proved by exploiting the properties of isotropic unit vector described in [45]. Lemma 2.2 can be proved by the definition of chi

[1] It is possible to use other schemes such as OMA or multi-user beamforming when the alignment is below the threshold. One such hybrid scheme was proposed in [38] and we will study the hybrid schemes in Chapter 3.

square distribution and the relationship between gamma distribution and chi square distribution. And Lemma 2.3 is proved as Lemma 3 in [46].

With Criterion-I, the instantaneous required transmit power for the alignment-based NOMA scheme to guarantee the SINR levels for both users, denoted as $P_{n,I}^R$, is given by (2.8) and the following theorem is proved for the mean value of $P_{n,I}^R$.

**Theorem 2.1** *Define*

$$\tilde{P}_{lo} \triangleq \frac{\gamma_H(1+\gamma_T)}{(M-1)\beta_{H_I}} + \frac{\gamma_T}{(M-1)\beta_{T_I}\rho_{th}^2} F(1,1;M;1-\rho_{th}^{-2}), \tag{2.10}$$

*where $F(\cdot,\cdot;\cdot;\cdot)$ is the hypergeometric function [47] and $\rho_{th}$ is the alignment threshold. The average required transmit power for the alignment-based NOMA with Criterion-I to guarantee SINR levels of both users, $\gamma_H$ and $\gamma_T$, has the following lower and upper bounds:*

$$\tilde{P}_{lo} \leq \mathbb{E}\left[P_{n,I}^R\right] \leq \tilde{P}_{lo}\left(1+\min\left\{\frac{\beta_{T_I}}{\beta_{H_I}}, \frac{\gamma_T}{\gamma_H}\right\}\right). \tag{2.11}$$

*Proof:* From (2.8), by upper bounding the min-function with $\beta_{T_I}\|\boldsymbol{h}_{T_I}\|^2\rho^2$, a lower bound of the instantaneous required transmit power with Criterion-I can be obtained by:

$$P_{n,I}^R \geq P_{lo} \triangleq \frac{\gamma_H(1+\gamma_T)}{\beta_{H_I}\|\boldsymbol{h}_{H_I}\|^2} + \frac{\gamma_T}{\beta_{T_I}\|\boldsymbol{h}_{T_I}\|^2\rho^2}.$$

Thus, by using Lemma 2.1, 2.2 and 2.3, the following can be derived:

$$\begin{aligned}
\mathbb{E}\left[P_{n,I}^R\right] \geq & \iiint_{V'} f_{\|\boldsymbol{h}_{H_I}\|^2}(x) f_{\|\boldsymbol{h}_{T_I}\|^2}(y) f_{\rho^2|\rho^2\geq\rho_{th}^2}(z) P_{lo}\mathrm{d}x\mathrm{d}y\mathrm{d}z \\
= & \underbrace{\frac{\gamma_H(1+\gamma_T)}{(M-1)\beta_{H_I}}}_{T_0} + \underbrace{\iiint_{V'} f_{\|\boldsymbol{h}_{T_I}\|^2}(y) f_{\rho^2|\rho^2\geq\rho_{th}^2}(z)\frac{\gamma_T}{\beta_{T_I}yz}\mathrm{d}x\mathrm{d}y\mathrm{d}z}_{T} \qquad (2.12) \\
= & T_0 + \frac{\gamma_T}{\left(1-\rho_{th}^2\right)^{M-1}\beta_{T_I}}\int_{\rho_{th}^2}^1 \frac{\left(1-\rho^2\right)^{M-2}}{\rho^2}\mathrm{d}\rho^2 = \tilde{P}_{lo}, \qquad (2.13)
\end{aligned}$$

where $V' = \{(x,y,z)|x\in(0,\infty), y\in(0,\infty), z\in[\rho_{th}^2,1]\}$, $f_X(\cdot)$ represents the PDF of the random variable $X$ and $f_{\rho^2|\rho^2\geq\rho_{th}^2}(z)$ is the conditional PDF of $\rho^2$

given by

$$f_{\rho^2|\rho^2\geq\rho_{th}^2}(z)=\frac{f_{\rho^2}(z)}{\mathbb{P}\left[\rho^2\geq\rho_{th}^2\right]}=\frac{M-1}{\left(1-\rho_{th}^2\right)^{M-1}}(1-z)^{M-2}.$$

By considering the integral representation of the hypergeometric function given by

$$F\left(\alpha,\beta;\gamma;z\right)=\frac{1}{\mathrm{B}(\beta,\gamma-\beta)}\int_0^1 t^{\beta-1}(1-t)^{\gamma-\beta-1}(1-tz)^{-\alpha}\,\mathrm{d}t,$$

where $\mathrm{B}(\cdot,\cdot)$ is the beta function, the second term of (2.13) can be further calculated by

$$\begin{aligned}&\frac{\gamma_T}{\left(1-\rho_{th}^2\right)^{M-1}\beta_{T_I}}\int_{\rho_{th}^2}^1\frac{\left(1-\rho^2\right)^{M-2}}{\rho^2}\mathrm{d}\rho^2\\ \overset{(a)}{=}&\frac{\gamma_T}{(M-1)\beta_{T_I}\rho_{th}^2}\frac{1}{\mathrm{B}(1,M-1)}\int_0^1\frac{(1-y)^{M-2}}{1-\left(1-\rho_{th}^{-2}\right)y}\,\mathrm{d}y\\ =&\frac{\gamma_T}{(M-1)\beta_{T_I}\rho_{th}^2}F\left(1,1;M;1-\rho_{th}^{-2}\right),\end{aligned}$$

where $(a)$ is obtained by the change of variable $y=\frac{1}{1-\rho_{th}^2}\rho^2-\frac{\rho_{th}^2}{1-\rho_{th}^2}$ and (2.10) is proved.

Next, we show the upper bound. First, define $V_1\triangleq\{(x,y,z)|\beta_{H_I}x<\beta_{T_I}yz\}$ and $V_2\triangleq\{(x,y,z)|\beta_{H_I}x\geq\beta_{T_I}yz\}$. By noticing that $V'=V_1\cup V_2$, we have from (2.12),

$$\begin{aligned}\mathbb{E}\left[P_{n,I}^R\right]=&T_0+\underbrace{\iiint_{V_1}f_{\|\boldsymbol{h}_{H_I}\|^2}(x)\frac{\gamma_T}{\beta_{H_I}x}\mathrm{d}x\mathrm{d}y\mathrm{d}z}_{T_1}+\\ &\underbrace{\iiint_{V_2}f_{\|\boldsymbol{h}_{T_I}\|^2}(y)f_{\rho^2}(z)\frac{\gamma_T}{\beta_{T_I}yz}\mathrm{d}x\mathrm{d}y\mathrm{d}z}_{T_2}.\end{aligned}\tag{2.14}$$

Since $V_2\subseteq V'$, we have $T_2\leq T$ and thus

$$T_0+T_2\leq T_0+T=\tilde{P}_{lo}.\tag{2.15}$$

For $T_1$, it can be shown that

$$\begin{aligned} T_1 &\leq \frac{\gamma_T}{\gamma_H} \iiint_{V'} f_{\|\boldsymbol{h}_{H_I}\|^2}(x) \frac{\gamma_H}{\beta_{H_I} x} \mathrm{d}x\mathrm{d}y\mathrm{d}z \\ &\leq \frac{\gamma_T}{\gamma_H} \iiint_{V'} f_{\|\boldsymbol{h}_{H_I}\|^2}(x) \frac{\gamma_H(1+\gamma_T)}{\beta_{H_I} x} \mathrm{d}x\mathrm{d}y\mathrm{d}z = \frac{\gamma_T}{\gamma_H} T_0, \end{aligned} \tag{2.16}$$

$$\begin{aligned} T_1 &\leq \frac{\beta_{T_I}}{\beta_{H_I}} \iiint_{V'} f_{\|\boldsymbol{h}_{T_I}\|^2}(y) \frac{\gamma_T}{\beta_{T_I} y} \mathrm{d}x\mathrm{d}y\mathrm{d}z \\ &\leq \frac{\beta_{T_I}}{\beta_{H_I}} \iiint_{V'} f_{\|\boldsymbol{h}_{T_I}\|^2}(y) f_{\rho^2}(z) \frac{\gamma_T}{\beta_{T_I} yz} \mathrm{d}x\mathrm{d}y\mathrm{d}z = \frac{\beta_{T_I}}{\beta_{H_I}} T. \end{aligned} \tag{2.17}$$

From (2.16) and (2.17) we can obtain:

$$T_1 \leq \min\left\{\frac{\beta_{T_I}}{\beta_{H_I}}, \frac{\gamma_T}{\gamma_H}\right\}(T_0 + T) = \min\left\{\frac{\beta_{T_I}}{\beta_{H_I}}, \frac{\gamma_T}{\gamma_H}\right\} \tilde{P}_{lo}. \tag{2.18}$$

By combining (2.14), (2.15) and (2.18), the upper bound of $\mathbb{E}\left[P_{n,I}^R\right]$ in (2.11) is proved.

□

*Remark 1:* The the average required transmit power and its bounds in Theorem 2.1 are calculated only under the circumstance that $\rho^2 \geq \rho_{th}^2$. For the average required transmit power over all circumstances, both the bounds should be scaled with $\left(1-\rho_{th}^2\right)^{M-1}$, i.e.,

$$\left(1-\rho_{th}^2\right)^{M-1} \tilde{P}_{lo} \leq \mathbb{E}\left[P_{n,I}^{\mathrm{ALL}}\right] \leq \left(1-\rho_{th}^2\right)^{M-1} \tilde{P}_{lo} \left(1 + \min\left\{\frac{\beta_{T_I}}{\beta_{H_I}}, \frac{\gamma_T}{\gamma_H}\right\}\right).$$

Theorem 2.1 provides a lower and an upper bound on the average required transit power. It is the foundation of subsequent analysis. In what follows, we provide several corollaries based on Theorem 2.1 for more insights.

**Corollary 2.1** *For any $\gamma_H > 0, \gamma_T > 0$, $\mathbb{E}[P_{n,I}^R]$ is unbounded when $\rho_{th} = 0$, in other words, $\lim_{\rho_{th}\to 0} \mathbb{E}[P_{n,I}^R] = \infty$. When $\rho_{th} > 0$, $\mathbb{E}[P_{n,I}^R]$ is bounded, in other words, $\mathbb{E}[P_{n,I}^R] < \infty$.*

*Proof:* Considering the alternative form of $\tilde{P}_{lo}$ given by (2.13), we have:

$$\mathbb{E}\left[P_{n,I}^R\right] \geq \tilde{P}_{lo} \geq \frac{\gamma_T}{\left(1-\rho_{th}^2\right)^{M-1} \beta_{T_I}} \int_{\rho_{th}^2}^{1} \frac{\left(1-\rho^2\right)^{M-2}}{\rho^2} \mathrm{d}\rho^2. \tag{2.19}$$

When $\rho_{th} \to 0$, by using binomial theorem, the limit of the integral in (2.19) can be calculated as

$$\begin{aligned}
\lim_{\rho_{th}\to 0} \int_{\rho_{th}^2}^{1} \frac{\left(1-\rho^2\right)^{M-2}}{\rho^2}\, \mathrm{d}\rho^2 &= \lim_{\rho_{th}\to 0} \int_{\rho_{th}^2}^{1} \sum_{k=0}^{M-2} \binom{M-2}{k} (-1)^k \left(\rho^2\right)^{k-1}\, \mathrm{d}\rho^2 \\
&= \sum_{k=1}^{M-2} \frac{1}{k} \binom{M-2}{k} (-1)^k \left(\rho^2\right)^k + \lim_{\rho_{th}^2\to 0} \ln\left|\rho^2\right| \Big|_{\rho_{th}^2}^{1} \\
&= \sum_{k=1}^{M-2} \frac{1}{k} \binom{M-2}{k} (-1)^k - \lim_{\rho_{th}^2\to 0} \ln \rho_{th}^2 \\
&= \infty.
\end{aligned} \tag{2.20}$$

By substituting (2.20) into (2.19), the average required transmit power is proved to be unbounded when $\rho_{th} = 0$.

For any $\rho_{th} > 0$, we have:

$$\int_{\rho_{th}^2}^{1} \frac{\left(1-\rho^2\right)^{M-2}}{\rho^2}\, \mathrm{d}\rho^2 = \sum_{k=1}^{M-2} \frac{1}{k} \binom{M-2}{k} (-1)^k - \ln \rho_{th}^2 < \infty,$$

thus $\tilde{P}_{lo} < \infty$. Therefore, from the upper bound of $\mathbb{E}\left[P_{n,I}^R\right]$ we have:

$$\mathbb{E}\left[P_{n,I}^R\right] \leq \tilde{P}_{lo} \left(1 + \min\left\{\frac{\beta_{T_I}}{\beta_{H_I}}, \frac{\gamma_T}{\gamma_H}\right\}\right) < \infty,$$

which concludes the proof. □

The results in Corollary 2.1 mean that if the alignment threshold is 0, i.e., the original NOMA is used, any user SINR constraints cannot be guaranteed with finite average transmit power. This is problematic in energy efficiency. Equation (2.20) shows that the unbounded average required transmit power is caused by the scenario when $\rho$ is in the vicinity of zero, i.e., the channel vectors are close-to-orthogonal. Naturally, as shown in Fig 2.1 and (2.8), a beam cannot serve the user whose channel orthogonal channel to the beam. More specifically, this user is the tail user as the beamformer is based on the channel of the head user. Thus the SINR constraint $\gamma_T$ cannot be achieved. This motivates the alignment-based NOMA scheme with a non-zero alignment threshold.

**Corollary 2.2** *When $\gamma_H \gg \gamma_T$ or $\beta_{H_I} \gg \beta_{T_I}$, the average required transmit power of the alignment-based NOMA can be tightly approximated as $\tilde{P}_{lo}$, i.e., $\mathbb{E}[P_{n,I}^R] \approx \tilde{P}_{lo}$.*

The result in Corollary 2.2 can be obtained directly from (2.11). A typical application scenario for NOMA is when one user has a significantly stronger channel and a large difference on the large-scale fading, i.e., $\beta_{H_I} \gg \beta_{T_I}$ is beneficial [33]. Given difference in the channel gains, it is also reasonable to expect a significantly better service to the stronger user, i.e., $\gamma_H \gg \gamma_T$. For these scenarios, Corollary 2.2 provides a tight approximation on the average required transmit power for any SINR constraints. The approximation $\tilde{P}_{lo}$ in (2.10) shows the behaviour of the average required transmit power with respect to the network parameters. For example, $\tilde{P}_{lo}$ increases with $\gamma_H$ and $\gamma_T$ while decreases with $\rho_{th}$, $\beta_{H_I}$, $\beta_{T_I}$ and $M$.

To further explore the behavior of $\mathbb{E}[P_{n,I}^R]$ with respect to $M$ and $\rho_{th}$, we introduce the following asymptotic result.

**Corollary 2.3** *When $\gamma_H \gg \gamma_T$ or $\beta_{H_I} \gg \beta_{T_I}$, for any fixed $M$, when $\rho_{th} \to 0$,*

$$\mathbb{E}[P_{n,I}^R] \approx \frac{\gamma_H(1+\gamma_T)}{(M-1)\beta_{H_I}} - \frac{\gamma_T}{\beta_{T_I}} \frac{\ln \rho_{th}^2 + \psi(M-1) + C}{\left(1-\rho_{th}^2\right)^{M-1}}, \tag{2.21}$$

*where $\psi(\cdot)$ is the di-gamma function and $C \approx 0.5772$ is the Euler-Mascheroni constant.*

*Proof:* When $\gamma_H \gg \gamma_T$ or $\beta_{H_I} \gg \beta_{T_I}$, we have from Corollary 2.2 $\mathbb{E}[P_{n,I}^R] \approx \tilde{P}_{lo}$ as defined in (2.10). Consider the alternative form of $\tilde{P}_{lo}$ in (2.13). The integral in (2.13) can be further calculated as follows:

$$\begin{aligned}
&\int_{\rho_{th}^2}^1 \frac{\left(1-\rho^2\right)^{M-2}}{\rho^2} \mathrm{d}\rho^2 \\
&= \sum_{k=0}^{M-2} (-1)^k \int_{\rho_{th}^2}^1 \binom{M-2}{k} \rho^{2(k-1)} d\rho^2 \\
&= \sum_{k=1}^{M-2} \frac{(-1)^k}{k} \binom{M-2}{k} \left(1-\rho_{th}^{2k}\right) - \ln \rho_{th}^2 \\
&= -\psi(M-1) - C - \ln \rho_{th}^2 + \mathcal{O}(\rho_{th}^2).
\end{aligned} \tag{2.22}$$

When $\rho_{th}^2 \to 0$, by ignoring the higher order terms of $\rho_{th}^2$, (2.21) is obtained.

□

This corollary provides a closed-form expression for the average required transmit power when $\rho_{th}$ is close to zero. It shows that the average required transmit power increases as $\ln(1/\rho_{th}^2)$ for small threshold.

With the proposed NOMA scheme, the users are in outage if and only if the BS is silent, i.e., the alignment of channel directions is smaller than the threshold. The outage probability is thus,

$$P_{\text{out}} = \mathbb{P}[\rho < \rho_{th}] = 1 - \left(1 - \rho_{th}^2\right)^{M-1}, \tag{2.23}$$

which is an increasing function of $M$ and $\rho_{th}$. With a fixed $M$, the outage probability increases as $\rho_{th}$ increases; at the same time, the power consumption decreases. Thus we can adjust the balance between power consumption and outage performance via the design of $\rho_{th}$. For the original NOMA, Corollary 2.1 shows that the average required transmit power is unbounded and (2.23) shows that the outage probability is zero. Thus the original NOMA achieves one of the singular end-point on the power-outage tradeoff curve, while the proposed NOMA scheme provides access to any point on the tradeoff curve by adjusting the alignment threshold.

Next, we consider massive BS antenna scenario where $M \gg 1$ and study the asymptotic behaviour and the scaling laws of the transmit power and the power-outage tradeoff. Notice that with a fixed $\rho_{th}$ value, the outage probability increases as $M$ increases, though the average required transmit power reduces. Thus for systems with massive BS antennas, a threshold design where $\rho_{th}$ decreases with $M$ is desirable. For this matter, we have the following results.

**Corollary 2.4** *When $M \to \infty$ and $\rho_{th}^2 = \lambda/M^\tau$ for a constant $\lambda > 0$, the following results on the average required transmit power and the outage probability can be obtained:*

- *when $\tau > 1$, $P_{\text{out}} \to 0$ and $\tilde{P}_{lo} \to \infty$;*
- *when $\tau < 1$, $P_{\text{out}} \to 1$ and $\tilde{P}_{lo} \to 0$;*

- *when* $\tau = 1$, $P_{\text{out}} \to 1 - e^{-\lambda}$ *and* $\tilde{P}_{lo} \to \frac{\gamma_T}{\beta_{T_I}} e^{\lambda} E_1(\lambda)$,

*where* $E_1(\cdot)$ *is the exponential integral function.*

*Proof:* The limits of $P_{out}$ in Corollary 2.4 can be obtained by:

$$\lim_{M\to\infty} \left(1 - \frac{\lambda}{M^\tau}\right)^{M-1} = \begin{cases} 0 & ,\tau < 1 \\ e^{-\lambda} & ,\tau = 1 \\ 1 & ,\tau > 1. \end{cases}$$

For the average transmit power, we consider the alternative form of $\tilde{P}_{lo}$ given by (2.13). When $\rho_{th}^2 = \lambda/M^\tau$, the limit of $\tilde{P}_{lo}$ is given by:

$$\lim_{M\to\infty} \tilde{P}_{lo} = \lim_{M\to\infty} \left[ \frac{\gamma_T}{\left(1 - \frac{\lambda}{M^\tau}\right)^{M-1} \beta_{T_I}} I \right],$$

where

$$I \triangleq \int_{\frac{\lambda}{M^\tau}}^{1} \frac{(1-\rho^2)^M}{\rho^2} \mathrm{d}\rho^2 \overset{y=M\rho^2}{=\!=\!=} \int_{\lambda M^{1-\tau}}^{M} \frac{\left(1 - \frac{y}{M}\right)^M}{y} \mathrm{d}y.$$

For any $y > 0$, since $\left(1 + \frac{y}{M}\right)^M$ increases with $M$ and converges to $e^y$ and $\left(1 + \frac{y}{M}\right)^{M+1}$ decreases with $M$ and converges to $e^y$ when $M \to \infty$, we have

$$\left(1 + \frac{y}{M}\right)^M < e^y < \left(1 + \frac{y}{M}\right)^{M+1} \quad \text{when } y > 0, M \geq 1. \tag{2.24}$$

Thus

$$\left(1 - \frac{y}{M}\right)^M e^y \geq \left(1 - \frac{y}{M}\right)^M \left(1 + \frac{y}{M}\right)^M = \left(1 - \frac{y^2}{M^2}\right)^M,$$

from which we have

$$\left(1 - \frac{y}{M}\right)^M \geq \left(1 - \frac{y^2}{M^2}\right)^M e^{-y}.$$

Further from (2.24),

$$\left(1 - \frac{y}{M}\right)^M e^{\frac{M}{M+1}y} < \left(1 - \frac{y^2}{M^2}\right)^M < 1,$$

from which we have

$$\left(1 - \frac{y}{M}\right)^M < e^{-\frac{M}{M+1}y} \leq e^{-\frac{1}{2}y}.$$

These gives, for all $M \geqslant 1$,

$$\frac{(1 - \frac{y^2}{M^2})^M}{y} e^{-y} \leqslant \frac{(1 - \frac{y}{M})^M}{y} \leqslant \frac{1}{y} e^{-\frac{y}{2}}. \tag{2.25}$$

From (2.25), an upper bound of $I$ is obtained as follows.

$$\begin{aligned} I &\geq \int_{\lambda M^{1-\tau}}^{M} \frac{\left(1-\frac{y^2}{M^2}\right)^M e^{-y}}{y} \mathrm{d}y \\ &\geq \left(1-\frac{1}{M}\right)^M \int_{\lambda M^{1-\tau}}^{M} \frac{e^{-y}}{y} \mathrm{d}y = \left(1-\frac{1}{M}\right)^M \left[E_1\left(\lambda M^{1-\tau}\right) - E_1(M)\right]. \end{aligned}$$

When $\tau > 1$, we have

$$\lim_{M\to\infty} I \geq \lim_{M\to\infty} \left(1-\frac{1}{M}\right)^M \left[E_1\left(\lambda M^{1-\tau}\right) - E_1(M)\right] = \infty,$$

and

$$\lim_{M\to\infty} \tilde{P}_{lo} = \infty.$$

When $\tau = 1$, by (2.25) and $\left(1-\frac{y}{M}\right)^M \to e^{-y}$, we have

$$I \to \int_{\lambda}^{\infty} y^{-1} e^{-y} \mathrm{d}y$$

through Lebesgue's dominated convergence theorem and the limit of $\tilde{P}_{lo}$ is given by

$$\lim_{M\to\infty} \tilde{P}_{lo} = \frac{\gamma_T}{\beta_{T_I}} e^{\lambda} E_1(\lambda)$$

When $\tau < 1$ we have

$$\begin{aligned} \tilde{P}_{lo} =& \frac{1}{\left(1-\frac{\lambda}{M^\tau}\right)^M} \int_{\lambda M^{1-\tau}}^{M} \frac{\left(1-\frac{y}{M}\right)^M}{y} \mathrm{d}y \\ \overset{(b)}{=}& \frac{1}{\left(1-\frac{\lambda}{M^\tau}\right)^M} \int_{\lambda}^{M^\tau} \frac{\left(1-\frac{z}{M^\tau}\right)^M}{z} \mathrm{d}z \\ \leq& \frac{1}{\lambda} \int_{\lambda}^{M^\tau} \left(\frac{1-\frac{z}{M^\tau}}{1-\frac{\lambda}{M^\tau}}\right)^M \mathrm{d}z = \frac{M^\tau - \lambda}{\lambda(M+1)}, \end{aligned}$$

where $(b)$ is obtained by the change of variable $z = y/M^{1-\tau}$ and thus

$$0 \leq \lim_{M\to\infty} \tilde{P}_{lo} \leq \lim_{M\to\infty} \frac{M^\tau - \lambda}{\lambda(M+1)} = 0.$$

This proves the corollary.

□

Corollary 2.4 shows the limits of $\mathbb{E}[P_{n,I}^R]$ and $P_{out}$ when $M$ increases. The two performance measures naturally compete with each other since a higher

outage probability means less transmissions and less power consumption realized with a higher $\rho_{th}$. The most interesting threshold design is when $\tau = 1$, meaning that the square of the alignment threshold decreases linearly in $M$, i.e., $\rho_{th}^2 = \lambda/M$ for a fixed $\lambda$. In this case, both $\mathbb{E}[P_{n,I}^R]$ and $P_{out}$ have non-trivial bounded limits and by adjusting the value of $\lambda$, we can achieve a continuous tradeoff curve for the power consumption and outage probability. Another observation is that the limits are independent of $\gamma_H$ and $\beta_{H_I}$, the two parameters of the stronger user. The outage probability limit is also independent of the parameters of the weaker user, while the average power consumption depends on $\gamma_T$ and $\beta_{T_I}$, meaning that the power consumption of the alignment-based NOMA scheme in a massive BS antenna scenario is dominated by the weaker user.

## 2.4 Average required transmit power of the alignment-based NOMA with Criterion-II

With Criterion-II, the instantaneous required transmit power for the alignment-based NOMA scheme to guarantee the SINR levels for both users can be reduced from (2.8) to the following:

$$P_{n,II}^R \triangleq \frac{\gamma_H(1+\gamma_T)}{\beta_{H_{II}}\|\boldsymbol{h}_{H_{II}}\|^2} + \frac{\gamma_T}{\beta_{T_{II}}\|\boldsymbol{h}_{T_{II}}\|^2\rho^2}. \tag{2.26}$$

The following theorem is proved for the mean value of $P_{n,II}^R$.

**Theorem 2.2** *Define*

$$A \triangleq \frac{\Gamma(2M-1)\left(\beta_{H_{II}}\beta_{T_{II}}\right)^{M-1}}{\Gamma(M)\Gamma(M+1)\left(\beta_{H_{II}}+\beta_{T_{II}}\right)^{2M-1}} \times \left[F\left(1,2M-1;M+1;\frac{\beta_{T_{II}}}{\beta_{H_{II}}+\beta_{T_{II}}}\right) + F\left(1,2M-1;M+1;\frac{\beta_{H_{II}}}{\beta_{H_{II}}+\beta_{T_{II}}}\right)\right]. \tag{2.27}$$

*The average required transmit power for the alignment-based NOMA scheme with Criterion-II to guarantee the SINR levels of both users, $\gamma_H$ and $\gamma_T$, is*

*given by:*

$$\begin{aligned}\mathbb{E}\left[P_{n,II}^{R}\right] =&\gamma_H\left(1+\gamma_T\right)A+\left[\beta_{H_{II}}^{-1}+\beta_{T_{II}}^{-1}-(M-1)A\right]\\ &\times\frac{\gamma_T}{(M-1)\rho_{th}^2}F\left(1,1;M;1-\rho_{th}^{-2}\right).\end{aligned}\tag{2.28}$$

*Proof:* Recall that $\|\boldsymbol{h}_1\|^2$ and $\|\boldsymbol{h}_2\|^2$ follow $\Gamma(M,1)$, the Gamma distribution with parameters $M$ and 1; $\rho^2$ follows Beta$(1,M-1)$, the Beta distribution with parameters 1 and $M-1$. Then we have $\beta_1\|\boldsymbol{h}_1\|^2\sim\Gamma\left(M,\beta_1\right)$ and $\beta_2\|\boldsymbol{h}_2\|^2\sim\Gamma\left(M,\beta_2\right)$. According to the definitions of head user and tail user in Criterion-II, we have

$$\begin{aligned}\alpha_H&\triangleq\beta_{H_{II}}\|\boldsymbol{h}_{H_{II}}\|^2=\max\left(\beta_1\|\boldsymbol{h}_1\|^2,\beta_2\|\boldsymbol{h}_2\|^2\right),\\ \alpha_T&\triangleq\beta_{T_{II}}\|\boldsymbol{h}_{T_{II}}\|^2=\min\left(\beta_1\|\boldsymbol{h}_1\|^2,\beta_2\|\boldsymbol{h}_2\|^2\right).\end{aligned}$$

For a random variable $X$, denote its cumulative distribution function (CDF) as $F_X(\cdot)$. Then, the CDF of $\alpha_H$ can be obtained by:

$$\begin{aligned}F_{\alpha_H}\left(x\right)=&\mathbb{P}\left[\max\left(\beta_1\|\boldsymbol{h}_1\|^2,\beta_2\|\boldsymbol{h}_2\|^2\right)\leq x\right]\\ =&F_{\beta_1\|\boldsymbol{h}_1\|^2}\left(x\right)\cdot F_{\beta_2\|\boldsymbol{h}_2\|^2}\left(x\right),\end{aligned}$$

and we can obtain the PDF of $\alpha_H$ by taking the derivative of $F_{\alpha_H}\left(x\right)$, which is given by:

$$\begin{aligned}f_{\alpha_H}\left(x\right)=&F_{\beta_1\|\boldsymbol{h}_1\|^2}\left(x\right)f_{\beta_2\|\boldsymbol{h}_2\|^2}\left(x\right)+f_{\beta_1\|\boldsymbol{h}_1\|^2}\left(x\right)F_{\beta_2\|\boldsymbol{h}_2\|^2}\left(x\right)\\ =&\frac{x^{M-1}e^{-\frac{x}{\beta_2}}}{\Gamma^2(M)\beta_2}\gamma\left(M,\frac{x}{\beta_1}\right)+\frac{x^{M-1}e^{-\frac{x}{\beta_1}}}{\Gamma^2(M)\beta_1}\gamma\left(M,\frac{x}{\beta_2}\right),\end{aligned}$$

where $\gamma(\cdot,\cdot)$ is the lower incomplete gamma function.

Similarly, the CDF and PDF of $\alpha_T$ can be obtained by:

$$F_{\alpha_T}\left(x\right)=F_{\beta_1\|\boldsymbol{h}_1\|^2}\left(x\right)+F_{\beta_2\|\boldsymbol{h}_2\|^2}\left(x\right)-F_{\beta_1\|\boldsymbol{h}_1\|^2}\left(x\right)\cdot F_{\beta_2\|\boldsymbol{h}_2\|^2}\left(x\right),$$

and

$$\begin{aligned}f_{\alpha_T}\left(x\right)=&f_{\beta_1\|\boldsymbol{h}_1\|^2}\left(x\right)+f_{\beta_2\|\boldsymbol{h}_2\|^2}\left(x\right)-f_{\alpha_H}(x)\\ =&\frac{x^{M-1}e^{-\frac{x}{\beta_1}}}{\Gamma(M)\beta_1}+\frac{x^{M-1}e^{-\frac{x}{\beta_2}}}{\Gamma(M)\beta_2}-f_{\alpha_H}(x).\end{aligned}$$

Then the average required transmit power can be calculated by the following integral:

$$\begin{aligned}\mathbb{E}\left[P_{n,II}^{R}\right] =& \iiint_{V} f_{\alpha_H}\left(\alpha_H\right) f_{\alpha_T}\left(\alpha_T\right) f_{\rho^2|\rho^2\geq\rho_{th}^2}\left(\rho^2\right) P_{\min}\mathrm{d}\alpha_H\mathrm{d}\alpha_T\mathrm{d}\rho^2 \\ =&\gamma_H(1+\gamma_T)\int_0^{\infty}\frac{f_{\alpha_H(x)}}{x}dx+\gamma_T\left(\int_0^{\infty}\frac{f_{\alpha_T(x)}}{x}dx\right)\left(\int_{\rho_{th}^2}^{1}\frac{f_{\rho^2}(x)}{\left(1-\rho_{th}^2\right)x}dx\right) \\ =&\gamma_H\left(1+\gamma_T\right)A+\frac{\frac{\gamma_T}{\beta_{H_{II}}}+\frac{\gamma_T}{\beta_{T_{II}}}-(M-1)\gamma_2 A}{\left(1-\rho_{th}^2\right)^{M-1}}\int_{\rho_{th}^2}^{1}\frac{\left(1-x\right)^{M-2}}{x}\mathrm{d}x.\end{aligned} \tag{2.29}$$

With some straightforward calculations and by using the definition of the Hypergeometric function, (2.28) can be obtained. □

Theorem 2.2 provides an exact expression for the average required transmit power with respect to the BS antenna number $M$, alignment threshold $\rho_{th}$, large-scale channel coefficients $\beta_{H_{II}}, \beta_{T_{II}}$, and SINR constraints $\gamma_H, \gamma_T$. And $A$ defined in (2.27) is the mean value of $1/\beta_{H_{II}}\|\boldsymbol{h}_{H_{II}}\|^2$, i.e.,

$$A=\int_0^{\infty}\frac{f_{\alpha_H(x)}}{x}dx.$$

Similarly to the analysis for the alignment-based NOMA with Criterion-I, for more insightful observations, we analyze the asymptotic behavior of the average power consumption and the results are given in the following corollaries.

**Corollary 2.5** *For any $\gamma_H > 0, \gamma_T > 0$, $\mathbb{E}[P_{n,II}^R]$ is unbounded when $\rho_{th} = 0$, in other words, $\lim_{\rho_{th}\to 0}\mathbb{E}[P_{n,II}^R] = \infty$. When $\rho_{th} > 0$, $\mathbb{E}[P_{n,II}^R]$ is bounded, in other words, $\mathbb{E}[P_{n,II}^R] < \infty$.*

*Proof:* By considering the integral form of $\mathbb{E}\left[P_{n,II}^{R}\right]$ given by (2.29), we have

$$\mathbb{E}\left[P_{n,II}^{R}\right]\geq\left[\frac{\gamma_T}{\beta_{H_{II}}}+\frac{\gamma_T}{\beta_{T_{II}}}-(M-1)\gamma_2 A\right]\frac{1}{\left(1-\rho_{th}^2\right)^{M-1}}\int_{\rho_{th}^2}^{1}\frac{\left(1-x\right)^{M-2}}{x}\mathrm{d}x.$$

When $\rho_{th}\to 0$, we have shown in (2.20) that

$$\lim_{\rho_{th}\to 0}\int_{\rho_{th}^2}^{1}\frac{\left(1-x\right)^{M-2}}{x}\mathrm{d}x=\infty.$$

Therefore,

$$\lim_{\rho_{th}\to 0} \mathbb{E}\left[P_{n,II}^{R}\right] = \infty.$$

For any $\rho_{th} > 0$, since $P_{n,II}^{R}$ is bounded, its mean value must be bounded, which concludes the proof.

□

For the original NOMA scheme with no alignment threshold, i.e., $\rho_{th} = 0$, Corollary 2.5 shows that the average transmit power is unbounded, indicating that NOMA is power inefficient when the user channel vectors are close to orthogonal.

**Corollary 2.6** *The average required transmit power for the alignment-based NOMA with Criterion-II has the following asymptotic behavior.*

- *For fixed $M$, when $\rho_{th} \to 0$,*

$$\begin{aligned}\mathbb{E}\left[P_{n,II}^{R}\right] \approx & \gamma_H\left(1+\gamma_T\right)A + \left[\frac{\gamma_T}{\beta_{H_{II}}} + \frac{\gamma_T}{\beta_{T_{II}}} - (M-1)\gamma_T A\right] \\ & \times \frac{-\psi(M-1) - C - \ln\rho_{th}^2}{\left(1-\rho_{th}^2\right)^{M-1}}.\end{aligned} \tag{2.30}$$

- *For fixed $\rho_{th} > 0$, when $M \to \infty$*

$$\begin{aligned}\mathbb{E}\left[P_{n,II}^{R}\right] \leq & \left[\min\left(\beta_{H_{II}}^{-1}, \beta_{T_{II}}^{-1}\right)\gamma_H(1+\gamma_T) + (\beta_{H_{II}}^{-1} + \beta_{T_{II}}^{-1})\frac{\gamma_T}{\rho_{th}^2}\right]\frac{1}{M-1} \\ & + \mathcal{O}\left(\frac{1}{M^2}\right).\end{aligned} \tag{2.31}$$

- *When $M \to \infty$ and $\rho_{th}^2 = \tau/M$ for a fixed $\tau$,*

$$\begin{aligned}\max\left(\beta_{H_{II}}^{-1}, \beta_{T_{II}}^{-1}\right)\gamma_T e^{\tau} E_1(\tau) + \mathcal{O}\left(\frac{1}{M}\right) \leq & \mathbb{E}\left[P_{n,II}^{R}\right] \\ \leq & (\beta_{H_{II}}^{-1} + \beta_{T_{II}}^{-1})\gamma_T e^{\tau} E_1(\tau) + \mathcal{O}\left(\frac{1}{M}\right).\end{aligned} \tag{2.32}$$

*Proof:* We have proved in (2.22) that

$$\int_{\rho_{th}^2}^{1} \frac{(1-x)^{M-2}}{x}\,\mathrm{d}x = -\psi(M-1) - C - \ln\rho_{th}^2 + \mathcal{O}(\rho_{th}^2). \tag{2.33}$$

When $M$ is fixed and $\rho_{th} \to 0$, by substituting (2.33) into (2.29) and ignoring the higher order term of $\rho_{th}^2$, (2.30) can be obtained.

Since $A$ is the mean value of $1/\beta_{H_{II}}\|\boldsymbol{h}_{H_{II}}\|^2$, we have:

$$A \leq \min\left(\mathbb{E}\left[\frac{1}{\beta_1\|\boldsymbol{h}_1\|^2}\right], \mathbb{E}\left[\frac{1}{\beta_2\|\boldsymbol{h}_2\|^2}\right]\right) = \frac{\min\left(\beta_{H_{II}}^{-1}, \beta_{T_{II}}^{-1}\right)}{M-1},$$

thus the first term of (2.29) can be upper bounded as:

$$\gamma_H\left(1+\gamma_T\right)A \leq \min\left(\beta_{H_{II}}^{-1}, \beta_{T_{II}}^{-1}\right)\frac{\gamma_H\left(1+\gamma_T\right)}{M-1}. \tag{2.34}$$

For the second term of (2.29) we have:

$$\max\left(\beta_{H_{II}}^{-1}, \beta_{T_{II}}^{-1}\right)\gamma_T \leq \frac{\gamma_T}{\beta_{H_{II}}} + \frac{\gamma_T}{\beta_{T_{II}}} - (M-1)\gamma_T A \leq (\beta_{H_{II}}^{-1} + \beta_{T_{II}}^{-1})\gamma_T, \tag{2.35}$$

and

$$\begin{aligned}
&\frac{1}{\left(1-\rho_{th}^2\right)^{M-1}}\int_{\rho_{th}^2}^{1}\frac{(1-x)^{M-2}}{x}\mathrm{d}x \overset{y=\frac{1-x}{1-\rho_{th}^2}}{=\!=\!=} \int_0^1 \frac{y^{M-2}}{1-\left(1-\rho_{th}^2\right)y}\,\mathrm{d}y \\
=&\int_0^1 y^{M-2}\sum_{n=0}^{\infty}\left[\left(1-\rho_{th}^2\right)y\right]^n\,\mathrm{d}y \\
=&\sum_{n=0}^{\infty}\frac{\left(1-\rho_{th}^2\right)^n}{M-1+n} < \frac{1}{M-1}\sum_{n=0}^{\infty}\left(1-\rho_{th}^2\right)^n = \frac{1}{(M-1)\rho_{th}^2}.
\end{aligned} \tag{2.36}$$

Then, (2.31) can be obtained by combining (2.34), (2.35) and (2.36).

When $M \to \infty$ and $\rho_{th}^2 = \tau/M$ for a fixed $\tau$, for the first term of (2.29) we have:

$$\lim_{M\to\infty}\gamma_H\left(1+\gamma_T\right)A \leq \min\left(\beta_{H_{II}}^{-1}, \beta_{T_{II}}^{-1}\right)\lim_{M\to\infty}\frac{\gamma_H\left(1+\gamma_T\right)}{M-1} = 0. \tag{2.37}$$

For the second term we have:

$$\lim_{M\to\infty}\frac{1}{\left(1-\frac{\lambda}{M}\right)^{M-1}}\int_{\frac{\lambda}{M}}^{1}\frac{(1-x)^{M-2}}{x}\mathrm{d}x = e^{\lambda}E_1(\lambda). \tag{2.38}$$

And (2.32) can be obtained by combining (2.35), (2.37) and (2.38).

□

Corollary 2.6 provides the asymptotic behaviors of the average required transmit power of the alignment-based NOMA scheme with Cirterion-II. The

results in (2.30) shows that for small alignment threshold, the required transmit power of the NOMA scheme increases as $-\ln \rho_{th}$. Further, from (2.31), it can be concluded that when the BS antenna number is large, with fixed alignment threshold, the required transmit power decreases at least linearly with the antenna number. On the other hand, when the square of the alignment threshold is designed to decrease linearly with the BS antenna number, the transmit power of NOMA is bounded from both sides by constants. This constant is independent of $\gamma_H$, but only depends on $\gamma_T$ and the large-scale fading coefficients. As larger $\rho_{th}$ leads to less NOMA transmissions thus more outage, the result in (2.32) also provides intelligence in the threshold design to tradeoff outage and transmit power for NOMA.

## 2.5 Simulation results

In this section, simulation results are demonstrated to show the performance of the alignment-based NOMA scheme and to verify our analytical results.

### 2.5.1 The alignment-based NOMA with Criterion-I

We simulate the NOMA application scenario where one cell-interior user and one cell-edge user are served [4] since Criterion-I determines the decoding order by the large-scale fading coefficients only. By setting $\beta_1 = 0$ dB and $\beta_2 = -10$ dB, the large-scale fading for User 1 is normalized and the large-scale fading of User 2 has 10 dB degradation. Therefore, User 1 is chosen as the head user while the User 2 is chosen as the tail user. Further, $\gamma_H = 10$ dB and $\gamma_T = 0$ dB are chosen considering the channel conditions of the users.

In Fig. 2.3, the average required transmit power of the alignment-based NOMA scheme to guarantee the SINR requirements for the two users is shown as a function of $\rho_{th}$. It can be seen that the average transmit power decreases with $\rho_{th}$. Further, the figure also shows that $\tilde{P}_{lo}$ in (2.10) is an accurate approximation for all parameter values, while the result in (2.21) is tight for small $\rho_{th}$, e.g., when $\rho_{th} < 0.1$. We can also see that the average transmit power increases linearly in $\log(1/\rho_{th})$ as $\rho_{th}$ approaches 0.

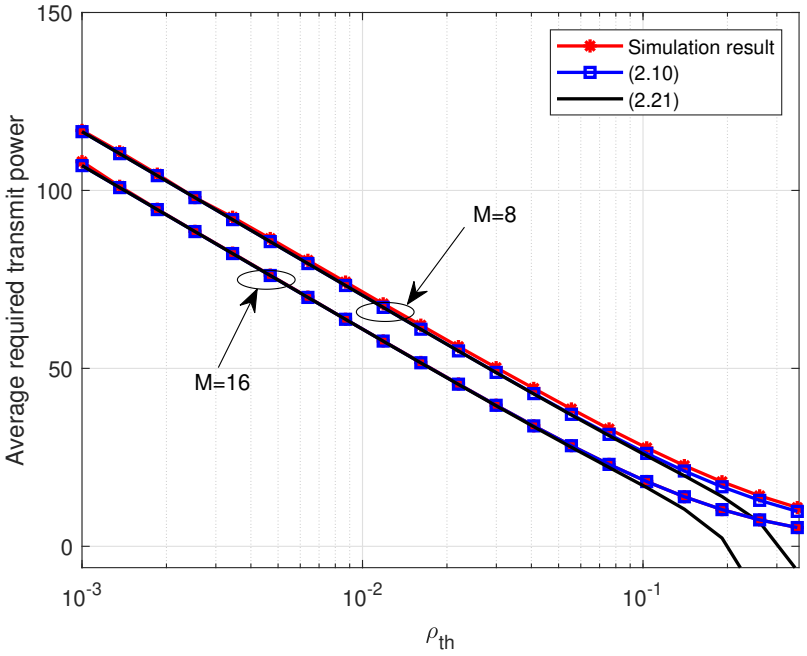


Figure 2.3: Average required transmit power versus $\rho_{th}$ where $M = 8$ and 16.

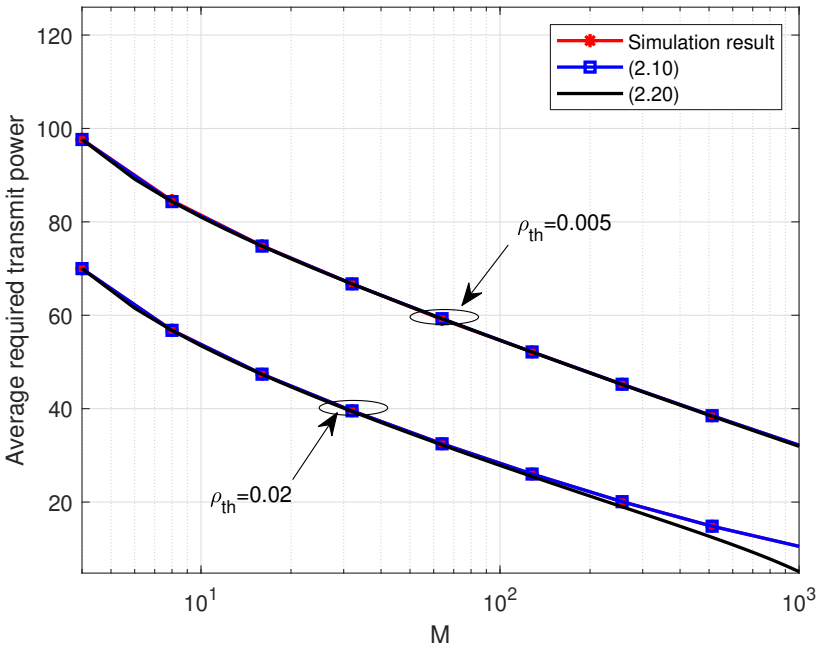


Figure 2.4: Average required transmit power versus $M$ where $\rho_{th} = 0.02$ and 0.005.

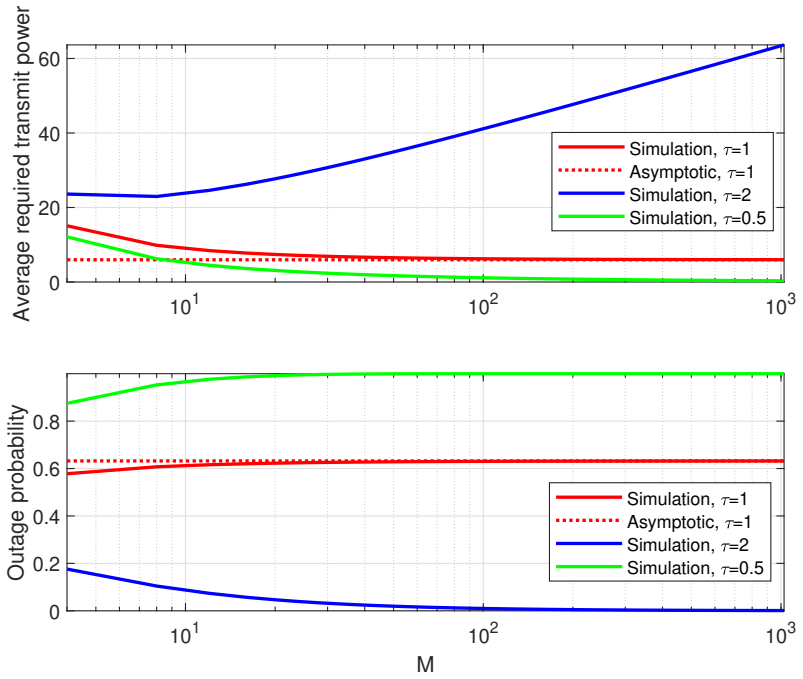


Figure 2.5: Average required transmit power and outage probability versus $M$ where $\rho_{th}^2 = 1/M^\tau$.

Fig. 2.4 depicts the average transmit power versus $M$. The accuracy of the approximations in (2.10) and (2.21) is verified again. The average required transmit power decreases as the number of antennas $M$ increases.

Fig. 2.5 shows the average required transmit power and outage probability versus $M$ where $\rho_{th}^2 = 1/M^\tau$. When $\tau = 1$, the average required transmit power decreases with $M$ and converges to a positive constant; when $\tau = 0.5 < 1$, it decreases with $M$ and approaches to 0; when $\tau = 2 > 1$, it increases unbounded with $M$, which validate the results in Corollary 2.4. And the behavior of the outage probability also matches the results in Corollary 2.4 in all three settings of $\tau$.

### 2.5.2 The alignment-based NOMA with Criterion-II

Criterion-II considers both large scale fading and small scale fading. When there is a large difference on the large scale fading coefficients of the two users, for example, 10dB difference, the performance is expected to be similar to that with Criterion-I, since the decoding order is dominated by the large scale

fading coefficients. Thus in this section we set $\beta_1 = \beta_2 = 0$dB, which is the normalized value, to have more dynamic decoding order based on the overall equivalent channel gain. Therefore, the decoding order is dynamic, and the user with larger equivalent channel gain will be chosen as the head user while another user will be chosen as the tail user. The values of $\gamma_H$ and $\gamma_T$ are the same as before consider the difference on equivalent channel gain.

In Fig. 2.6, the required average required transmit power to guarantee the SINR levels for both users is shown as a function of $\rho_{th}$ where $M = 8$ and 16. We compare the average power obtained by computer simulation, the exact analytical result given in (2.28), and the small-$\rho_{th}$ approximation obtained from (2.30) by ignoring higher order terms. We can find that the approximation is accurate with small $\rho_{th}$ (e.g., less than 0.1 for $M = 16$), while (2.28) has perfect match for all values of $\rho_{th}$. Another observation is that the average transmit power increases with $-\ln \rho_{th}$ for when $\rho_{th} \to 0$, which validates the asymptotic behavior claimed in Corollary 2.6.

Fig. 2.7 shows the average required transmit power versus $M$ with different threshold designs. Again, the correctness of Theorem 2.2 and the accuracy of the small-$\rho_{th}$ approximation are demonstrated. Further, it can also be seen that when $\rho_{th}$ has a fixed value, the average transmit power decreases with $M$; and for the large $M$ range, the decreases is linear in $1/M$. For the case of $\rho_{th}^2 = 0.01/M$, the transmit power approaches a constant as $M$ increases. These conform with our asymptotic results in Corollary 2.6.

## 2.6 Conclusion

A modified NOMA scheme based on the alignment of channel directions is proposed in this chapter for systems with a multiple-antenna BS and two single-antenna users. For this modified scheme with two criteria on deciding the decoding order, We derived the average required transmit power with SINR guarantee for both users and proved that a positive threshold is required for finite average transmit power. The results also show the behaviour of the average required transmit power with respect to the threshold and the BS

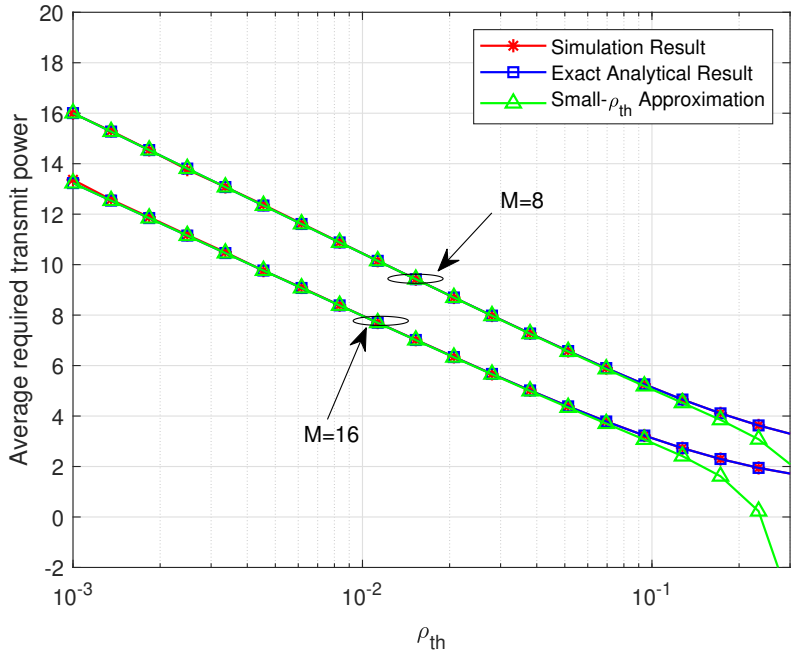


Figure 2.6: Average transmit power versus $\rho_{th}$ for a two-user cluster where $M = 8$ and 16, $\beta_1 = 0$dB, $\beta_2 = 0$dB, $\gamma_H = 10$dB, $\gamma_T = 0$dB.

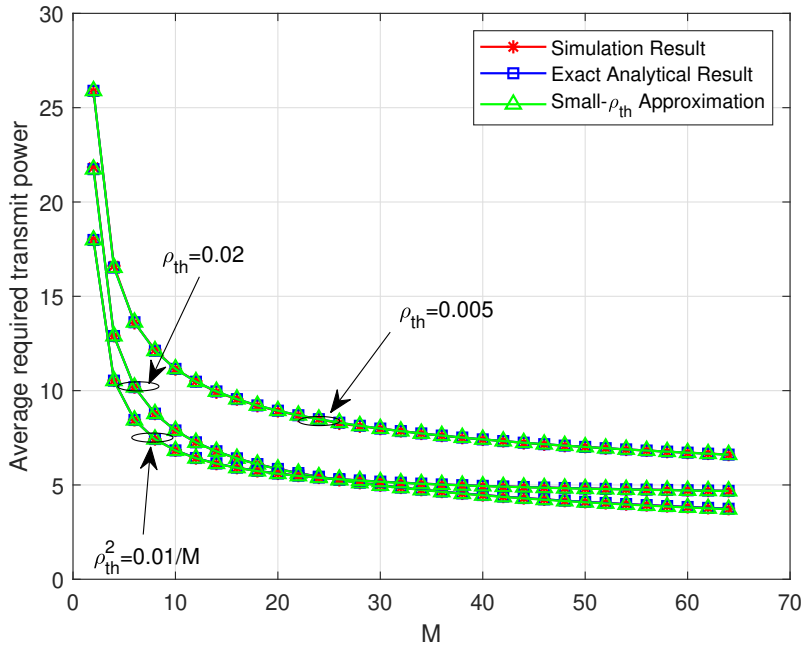


Figure 2.7: Average transmit power versus $M$ for a two-user cluster where $\beta_1 = 0$dB, $\beta_2 = 0$dB, $\gamma_H = 10$dB, $\gamma_T = 0$dB.

antenna number. Moreover, to balance the outage probability and average required transmit power in systems with massive BS antenna array, we proposed to design the threshold as a decreasing function of the number of BS antennas. The scaling laws of the outage probability and average required transmit power as well as their tradeoff law are obtained for different threshold designs.

# Chapter 3

# Power analysis of multi-user beamforming and hybrid design

In this chapter we firstly explore the power consumption of the multi-user beamforming scheme. By noticing that NOMA and multi-user beamforming have different preferred application scenarios, we propose an alignment-based hybrid scheme where NOMA transmission is conducted when $\rho$ is larger than a threshold and multi-user beamforming is used when $\rho$ is smaller than the threshold. This can utilize the advantages and avoid the disadvantages of the two schemes. Then, the power consumption of the hybrid scheme is studied with three alignment threshold selections.

The remainder of this chapter is organized as follow. We derive the instantaneous required transmit power for the multi-user beamforming scheme with SINR guarantee for both users in Section 3.1. Then, similar to the modified NOMA scheme, an alignment-based multi-user beamforming scheme is proposed to save power and the average required transmit power of this scheme is analyzed in Section 3.2. In Section 3.3, by combining the two alignment-based schemes, a hybrid scheme of NOMA and multi-user beamforming is obtained and several alignment threshold designs are proposed to reduce the average power consumption. Section 3.4 shows the simulation results on the power consumption of the alignment-based multi-user beamforming and the hybrid scheme while Section 3.5 is a summary of this chapter.

## 3.1 Instantaneous required transmit power for multi-user beamforming scheme with SINR guarantee for both users

We consider a network with one $M$-antenna BS serving two single-antenna users, and the channel model is described by (1.3) and (1.4). In the multi-user beamforming scheme, the BS serve the users with common time-frequency resource block but different beamformers while the users decode their own signals by treating the signals for the other user as noise.

We denote $\boldsymbol{b}_{1,m}$ and $\boldsymbol{b}_{2,m}$ as the beamformers for User 1 and User 2 and the subscript $m$ refers to the multi-user beamforming scheme. Then the transmit vector from the BS, $\boldsymbol{x}_m$ is given by:

$$\boldsymbol{x}_m = \sqrt{P_{1,m}}\boldsymbol{b}_{1,m}s_1 + \sqrt{P_{2,m}}\boldsymbol{b}_{2,m}s_2,$$

where $P_{1,m}, P_{2,m}$ are the power allocated to $s_1, s_2$ and the total transmit power is given by $P_m = P_{1,m} + P_{2,m}$. The received signal at the users can be written as:

$$\begin{aligned} y_{1,m} =& \sqrt{P_{1,m}}\boldsymbol{g}_1^H\boldsymbol{b}_{1,m}s_1 + \sqrt{P_{2,m}}\boldsymbol{g}_1^H\boldsymbol{b}_{2,m}s_2 + n_1, \\ y_{2,m} =& \sqrt{P_{1,m}}\boldsymbol{g}_2^H\boldsymbol{b}_{1,m}s_1 + \sqrt{P_{2,m}}\boldsymbol{g}_2^H\boldsymbol{b}_{2,m}s_2 + n_2. \end{aligned}$$

Since the two users decode their own signals by treating interference as noise, the SINRs for User $k$ to decode $s_k$ are given by:

$$\mathrm{SINR}_{1,m} = \frac{P_{1,m}\left|\boldsymbol{g}_1^H\boldsymbol{b}_{1,m}\right|^2}{1 + P_{2,m}\left|\boldsymbol{g}_1^H\boldsymbol{b}_{2,m}\right|^2}, \tag{3.1}$$

and

$$\mathrm{SINR}_{2,m} = \frac{P_{2,m}\left|\boldsymbol{g}_2^H\boldsymbol{b}_{2,m}\right|^2}{1 + P_{1,m}\left|\boldsymbol{g}_2^H\boldsymbol{b}_{1,m}\right|^2}. \tag{3.2}$$

By considering the MF beamforming with beamformers given by

$$\boldsymbol{b}_{k,m} = \frac{\boldsymbol{h}_k}{\|\boldsymbol{h}_k\|}, \quad k = 1, 2, \tag{3.3}$$

we can reduce (3.1) and (3.2) to

$$\mathrm{SINR}_{1,m} = \frac{P_{1,m}\beta_1\|\boldsymbol{h}_1\|^2}{1 + P_{2,m}\beta_1\|\boldsymbol{h}_1\|^2\rho^2}, \tag{3.4}$$

and

$$\mathrm{SINR}_{2,m} = \frac{P_{2,m}\beta_2\|\boldsymbol{h}_2\|^2}{1 + P_{1,m}\beta_2\|\boldsymbol{h}_2\|^2\rho^2}. \tag{3.5}$$

To guarantee the SINR requirements denoted as $\gamma_1$ and $\gamma_2$, we need

$$\mathrm{SINR}_{1,m} \geq \gamma_1, \tag{3.6}$$

$$\text{and} \qquad \mathrm{SINR}_{2,m} \geq \gamma_2. \tag{3.7}$$

By combining (3.4) and (3.6), we can get a lower bound of $P_{1,m}$

$$P_{1,m} \geq \frac{\gamma_1}{1+\gamma_1\rho^2}\left(\frac{1}{\beta_1\|\boldsymbol{h}_1\|^2} + P_m\rho^2\right). \tag{3.8}$$

And the lower bound of $P_{2,m}$ can be obtained by combining (3.5) and (3.7), which is given by:

$$P_{2,m} \geq \frac{\gamma_2}{\beta_2\|\boldsymbol{h}_2\|^2} + \frac{\gamma_1\gamma_2\rho^2}{1+\gamma_1\rho^2}\left(\frac{1}{\beta_1\|\boldsymbol{h}_1\|^2} + P_m\rho^2\right). \tag{3.9}$$

Then, by taking the summation of (3.8) and (3.9), with some calculations, we can obtain the instantaneous required transmit power for multi-user beamforming scheme to guarantee the SINR levels for both user as follow

$$P_m \geq P_m^R \triangleq \frac{\gamma_1 + \gamma_1\gamma_2\rho^2}{1-\gamma_1\gamma_2\rho^4}\frac{1}{\beta_1\|\boldsymbol{h}_1\|^2} + \frac{\gamma_2 + \gamma_1\gamma_2\rho^2}{1-\gamma_1\gamma_2\rho^4}\frac{1}{\beta_2\|\boldsymbol{h}_2\|^2}, \tag{3.10}$$

for the case of

$$\rho^4 < \frac{1}{\gamma_1\gamma_2}. \tag{3.11}$$

When the condition in (3.11) does not hold, any SINR requirement cannot be guaranteed.

It is not surprising that a large value of $\rho$ is detrimental to multi-user beamforming scheme since the users cause high interference with each other. A large value of $\rho$ means that their channels are close-to-parallel, i.e., the alignment of channel directions is high. This is also reflected by (3.4) and (3.5) where $P_{2,m}$ is in the interference component of $\mathrm{SINR}_{1,m}$ while $P_{1,m}$ is in the interference component of $\mathrm{SINR}_{2,m}$ and the interference part increase with $\rho$. Therefore, to increase either $P_{1,m}$ or $P_{2,m}$ is detrimental to another user as it increases the interference. Therefore, It is impossible to guarantee the SINR levels for all the users even with infinite power if (3.11) is not satisfied.

As a result, we can find that the instantaneous required transmit power $P_m^R$ increases with $\rho$ and approaches to infinity when $\rho^2$ approaches to $1/\sqrt{\gamma_1\gamma_2}$. Therefore, The condition on $\rho$ given by (3.11) is necessary while using the multi-user beamforming scheme and the users will be in outage when (3.11) is not satisfied.

## 3.2 Average transmit power for the alignment-based multi-user beamforming scheme with SINR guarantee for both users

As discussed in the last section, multi-user beamforming scheme cannot guarantee the user SINRs when $\rho$ is large. Therefore, similar to the alignment-based NOMA scheme, we propose an alignment-based multi-user beamforming scheme where the downlink data transmissions are conducted only when the alignment of channel directions is smaller than a threshold $\rho_{th}$ ($\rho_{th} < (\gamma_1\gamma_2)^{-1/4}$), otherwise the BS keeps silent to save power. And the following theorem is proved for the mean value of $P_m^R$ with this scheme.

**Theorem 3.1** *The average required transmit power for the alignment-based multi-user beamforming scheme to guarantee the SINR levels for both users, $\gamma_1$ and $\gamma_2$, is given by*

$$\mathbb{E}\left[P_m^R\right] = c_1 I_1 + c_2 I_2, \tag{3.12}$$

*where*

$$c_1 \triangleq \frac{1}{2}\left(\sqrt{\gamma_1+\gamma_2}\right)\left(\frac{\sqrt{\gamma_1}}{\beta_1}+\frac{\sqrt{\gamma_2}}{\beta_2}\right),$$

$$c_2 \triangleq \frac{1}{2}\left(\sqrt{\gamma_1-\gamma_2}\right)\left(\frac{\sqrt{\gamma_1}}{\beta_1}-\frac{\sqrt{\gamma_2}}{\beta_2}\right),$$

$$I_1 \triangleq (\gamma_1\gamma_2)^{-\frac{M-1}{2}}\left\{-\left(\sqrt{\gamma_1\gamma_2}-1\right)^{M-2}\ln\left(1-\sqrt{\gamma_1\gamma_2}\rho_{th}^2\right) + \sum_{k=1}^{M-2}\binom{M-2}{k}\left(\sqrt{\gamma_1\gamma_2}-1\right)^{M-2-k}\frac{1}{k}\left[1-\left(1-\sqrt{\gamma_1\gamma_2}\rho_{th}^2\right)^k\right]\right\},$$

$$I_2 \triangleq (\gamma_1\gamma_2)^{-\frac{M-1}{2}}\left\{\left(\sqrt{\gamma_1\gamma_2}+1\right)^{M-2}\ln\left(1+\sqrt{\gamma_1\gamma_2}\rho_{th}^2\right) + \sum_{k=1}^{M-2}\binom{M-2}{k}\left(\sqrt{\gamma_1\gamma_2}+1\right)^{M-2-k}\frac{(-1)^k}{k}\left[\left(1+\sqrt{\gamma_1\gamma_2}\rho_{th}^2\right)^k-1\right]\right\}.$$

*Proof:* By using Lemma 2.1, 2.2 and 2.3 in (3.10), the average required transmit power is given by

$$\begin{aligned}\mathbb{E}\left[P_m^R\right] =& \int_0^{\rho_{th}^2}\int_0^{\infty}\int_0^{\infty} P_m^R f_{\rho_{th}^2}(x) f_{\|\boldsymbol{h}_1\|^2}(y) f_{\|\boldsymbol{h}_2\|^2}(z)\mathrm{d}y\mathrm{d}z\mathrm{d}x \\ =& \frac{\gamma_1}{\beta_1}\int_0^{\rho_{th}^2}\frac{1+\gamma_2 x}{1-\gamma_1\gamma_2 x^2}(1-x)^{M-2}\mathrm{d}x + \frac{\gamma_2}{\beta_2}\int_0^{\rho_{th}^2}\frac{1+\gamma_1 x}{1-\gamma_1\gamma_2 x^2}(1-x)^{M-2}\mathrm{d}x \\ =& c_1\underbrace{\int_0^{\rho_{th}^2}\frac{(1-x)^{M-2}}{1-\sqrt{\gamma_1\gamma_2}x}\,\mathrm{d}x}_{I_1} + c_2\underbrace{\int_0^{\rho_{th}^2}\frac{(1-x)^{M-2}}{1+\sqrt{\gamma_1\gamma_2}x}\,\mathrm{d}x}_{I_2}.\end{aligned}$$

For $I_1$ we have

$$\begin{aligned} I_1 =& \int_0^{\rho_{th}^2}\frac{(1-x)^{M-2}}{1-\sqrt{\gamma_1\gamma_2}x}\,\mathrm{d}x \\ \overset{(c)}{=}& (\gamma_1\gamma_2)^{-\frac{M-1}{2}}\int_{1-\sqrt{\gamma_1\gamma_2}\rho_{th}^2}^{1} y^{-1}\left(\sqrt{\gamma_1\gamma_2}-1+y\right)^{M-2}\,\mathrm{d}y \\ =& (\gamma_1\gamma_2)^{-\frac{M-1}{2}}\left\{-\left(\sqrt{\gamma_1\gamma_2}-1\right)^{M-2}\ln\left(1-\sqrt{\gamma_1\gamma_2}\rho_{th}^2\right)\right. \\ &\left.+\sum_{k=1}^{M-2}\binom{M-2}{k}\left(\sqrt{\gamma_1\gamma_2}-1\right)^{M-2-k}\frac{1}{k}\left[1-\left(1-\sqrt{\gamma_1\gamma_2}\rho_{th}^2\right)^k\right]\right\}, \end{aligned} \tag{3.13}$$

where $(c)$ is obtained by the change of variable $y = 1-\sqrt{\gamma_1\gamma_2}x$. Similarly, for $I_2$ we have

$$\begin{aligned} I_2 =& \int_0^{\rho_{th}^2}\frac{(1-x)^{M-2}}{1+\sqrt{\gamma_1\gamma_2}x}\,\mathrm{d}x \\ \overset{(d)}{=}& (\gamma_1\gamma_2)^{-\frac{M-1}{2}}\int_1^{1+\sqrt{\gamma_1\gamma_2}\rho_{th}^2} y^{-1}\left(\underbrace{\sqrt{\gamma_1\gamma_2}+1}_{b}-y\right)^{M-2}\,\mathrm{d}y \\ =& (\gamma_1\gamma_2)^{-\frac{M-1}{2}}\left\{\left(\sqrt{\gamma_1\gamma_2}+1\right)^{M-2}\ln\left(1+\sqrt{\gamma_1\gamma_2}\rho_{th}^2\right)\right. \\ &\left.+\sum_{k=1}^{M-2}\binom{M-2}{k}\left(\sqrt{\gamma_1\gamma_2}+1\right)^{M-2-k}\frac{(-1)^k}{k}\left[\left(1+\sqrt{\gamma_1\gamma_2}\rho_{th}^2\right)^k-1\right]\right\}, \end{aligned} \tag{3.14}$$

where $(d)$ is obtained by the change of variable $y = 1+\sqrt{\gamma_1\gamma_2}x$. And Theorem 3.1 is proved. □

Theorem 3.1 provides an exact closed-form expression of the average required transmit power for the alignment-based multi-user beamforming scheme. For more insights, we provide the following corollaries.

**Corollary 3.1** *The average required transmit power* $\mathbb{E}\left[P_m^R\right]$ *is bounded by*

$$c_1 I_{1,L} + c_2 I_{2,L} \leq \mathbb{E}\left[P_m^R\right] \leq c_1 I_{1,U} + c_2 I_{2,U} \quad when \ \frac{\beta_2}{\sqrt{\gamma_2}} \geq \frac{\beta_1}{\sqrt{\gamma_1}} \tag{3.15}$$

*and*

$$c_1 I_{1,L} + c_2 I_{2,U} \leq \mathbb{E}\left[P_m^R\right] \leq c_1 I_{1,U} + c_2 I_{2,L} \quad when \ \frac{\beta_2}{\sqrt{\gamma_2}} < \frac{\beta_1}{\sqrt{\gamma_1}}, \tag{3.16}$$

*where*

$$I_{1,U} \triangleq \frac{1}{\sqrt{\gamma_1\gamma_2}} e^{-\frac{M-2}{\sqrt{\gamma_1\gamma_2}}} \left[E_1\left(-\frac{M-2}{\sqrt{\gamma_1\gamma_2}} + (M-2)\rho_{th}^2\right) - E_1\left(-\frac{M-2}{\sqrt{\gamma_1\gamma_2}}\right)\right],$$

$$I_{1,L} \triangleq \frac{1}{\sqrt{\gamma_1\gamma_2}} e^{-a\frac{M-2}{\sqrt{\gamma_1\gamma_2}}} \left[E_1\left(-a\frac{M-2}{\sqrt{\gamma_1\gamma_2}} + a(M-2)\rho_{th}^2\right) - E_1\left(-a\frac{M-2}{\sqrt{\gamma_1\gamma_2}}\right)\right],$$

$$I_{2,U} \triangleq \frac{1}{\sqrt{\gamma_1\gamma_2}} e^{\frac{M-2}{\sqrt{\gamma_1\gamma_2}}} \left[E_1\left(\frac{M-2}{\sqrt{\gamma_1\gamma_2}}\right) - E_1\left(\frac{M-2}{\sqrt{\gamma_1\gamma_2}} + (M-2)\rho_{th}^2\right)\right],$$

$$I_{2,L} \triangleq \frac{1}{\sqrt{\gamma_1\gamma_2}} e^{a\frac{M-2}{\sqrt{\gamma_1\gamma_2}}} \left[E_1\left(a\frac{M-2}{\sqrt{\gamma_1\gamma_2}}\right) - E_1\left(a\frac{M-2}{\sqrt{\gamma_1\gamma_2}} + a(M-2)\rho_{th}^2\right)\right],$$

$$a = -\frac{\ln\left(1-\rho_{th}^2\right)}{\rho_{th}^2}.$$

*Proof:* First, we introduce the following inequality

$$e^{-a(M-2)x} \leq (1-x)^{M-2} \leq e^{-(M-2)x}, \quad \text{when } 0 \leq x \leq \rho_{th}^2.$$

Then, an upper bound of $I_1$ can be obtained by

$$\begin{aligned} I_1 \leq I_{1,U} =& \int_0^{\rho_{th}^2} \frac{e^{-(M-2)x}}{1-\sqrt{\gamma_1\gamma_2}x}\, \mathrm{d}x \\ =& \frac{1}{\sqrt{\gamma_1\gamma_2}} e^{-\frac{M-2}{\sqrt{\gamma_1\gamma_2}}} \left[E_1\left(-\frac{M-2}{\sqrt{\gamma_1\gamma_2}} + (M-2)\rho_{th}^2\right) - E_1\left(-\frac{M-2}{\sqrt{\gamma_1\gamma_2}}\right)\right], \end{aligned}$$

and a lower bound of $I_1$ can be obtained by

$$\begin{aligned} I_1 \geq I_{1,L} =& \int_0^{\rho_{th}^2} \frac{e^{-a(M-2)x}}{1-\sqrt{\gamma_1\gamma_2}x}\, \mathrm{d}x \\ =& \frac{1}{\sqrt{\gamma_1\gamma_2}} e^{-a\frac{M-2}{\sqrt{\gamma_1\gamma_2}}} \left[E_1\left(-a\frac{M-2}{\sqrt{\gamma_1\gamma_2}} + a(M-2)\rho_{th}^2\right) - E_1\left(-a\frac{M-2}{\sqrt{\gamma_1\gamma_2}}\right)\right]. \end{aligned}$$

Similarly, for $I_2$ we have

$$I_2 \le I_{2,U} = \int_0^{\rho_{th}^2} \frac{e^{-(M-2)x}}{1+\sqrt{\gamma_1\gamma_2}x}\,\mathrm{d}x$$
$$= \frac{1}{\sqrt{\gamma_1\gamma_2}} e^{\frac{M-2}{\sqrt{\gamma_1\gamma_2}}} \left[ E_1\left(\frac{M-2}{\sqrt{\gamma_1\gamma_2}}\right) - E_1\left(\frac{M-2}{\sqrt{\gamma_1\gamma_2}} + (M-2)\rho_{th}^2\right)\right],$$

and

$$I_2 \ge I_{2,L} = \int_0^{\rho_{th}^2} \frac{e^{-a(M-2)x}}{1+\sqrt{\gamma_1\gamma_2}x}\,\mathrm{d}x$$
$$= \frac{1}{\sqrt{\gamma_1\gamma_2}} e^{a\frac{M-2}{\sqrt{\gamma_1\gamma_2}}} \left[ E_1\left(a\frac{M-2}{\sqrt{\gamma_1\gamma_2}}\right) - E_1\left(a\frac{M-2}{\sqrt{\gamma_1\gamma_2}} + a(M-2)\rho_{th}^2\right)\right].$$

Since $c_1 > 0$ and

$$\begin{cases} c_2 \ge 0 & \text{when } \frac{\beta_2}{\sqrt{\gamma_2}} \ge \frac{\beta_1}{\sqrt{\gamma_1}} \\ c_2 < 0 & \text{when } \frac{\beta_2}{\sqrt{\gamma_2}} < \frac{\beta_1}{\sqrt{\gamma_1}}, \end{cases}$$

the upper bound and lower bound of $\mathbb{E}\left[P_m^R\right]$ is proved. □

**Corollary 3.2** *When $\rho_{th}$ is fixed and $M$ is large, the average required transmit power for the alignment-based multi-user beamforming scheme to guarantee the SINR levels for both users decreases in $1/M$.*

*Proof:* By considering the integral form of $I_1$ in the first step of (3.13), an upper bound of $I_1$ is given as follows.

$$I_1 = \int_0^{\rho_{th}^2} \frac{(1-x)^{M-2}}{1-\sqrt{\gamma_1\gamma_2}x}\,\mathrm{d}x \le \underbrace{\int_0^{\rho_{th}^2} \frac{(1-x)^{M-2}}{1-\sqrt{\gamma_1\gamma_2}\rho_{th}^2}\,\mathrm{d}x}_{I_{1,U1}}$$
$$= -\frac{1}{1-\sqrt{\gamma_1\gamma_2}\rho_{th}^2} \int_0^{\rho_{th}^2} (1-x)^{M-2}\,\mathrm{d}(1-x)$$
$$= -\frac{1}{1-\sqrt{\gamma_1\gamma_2}\rho_{th}^2} \left[ \frac{1}{M-1}(1-x)^{M-1}\Big|_0^{\rho_{th}^2} \right]$$
$$= \frac{1}{(M-1)\left(1-\sqrt{\gamma_1\gamma_2}\rho_{th}^2\right)} \left[1-\left(1-\rho_{th}^2\right)^{M-1}\right].$$

For a lower bound of $I_1$, we have

$$\begin{aligned} I_1 = \int_0^{\rho_{th}^2} \frac{(1-x)^{M-2}}{1-\sqrt{\gamma_1\gamma_2}x}\,\mathrm{d}x \geq & \underbrace{\int_0^{\rho_{th}^2}(1-x)^{M-2}\,\mathrm{d}x}_{I_{1,L1}} \\ = & -\int_0^{\rho_{th}^2}(1-x)^{M-2}\,\mathrm{d}(1-x) \\ = & -\frac{1}{M-1}(1-x)^{M-1}\Big|_0^{\rho_{th}^2} \\ = & \frac{1}{(M-1)}\left[1-\left(1-\rho_{th}^2\right)^{M-1}\right]. \end{aligned}$$

Similarly, for $I_2$ we have

$$\begin{aligned} I_2 = \int_0^{\rho_{th}^2} \frac{(1-x)^{M-2}}{1+\sqrt{\gamma_1\gamma_2}x}\,\mathrm{d}x \leq & \underbrace{\int_0^{\rho_{th}^2}(1-x)^{M-2}\,\mathrm{d}x}_{I_{2,U1}} \\ = & \frac{1}{(M-1)}\left[1-\left(1-\rho_{th}^2\right)^{M-1}\right] \end{aligned}$$

and

$$\begin{aligned} I_2 = \int_0^{\rho_{th}^2} \frac{(1-x)^{M-2}}{1+\sqrt{\gamma_1\gamma_2}x}\,\mathrm{d}x \geq & \underbrace{\int_0^{\rho_{th}^2}\frac{(1-x)^{M-2}}{1+\sqrt{\gamma_1\gamma_2}x}\,\mathrm{d}x}_{I_{2,L1}} \\ = & \frac{1}{(M-1)\left(1+\sqrt{\gamma_1\gamma_2}\rho_{th}^2\right)}\left[1-\left(1-\rho_{th}^2\right)^{M-1}\right]. \end{aligned}$$

When $c_2 \geq 0$, recall the following upper and lower bounds of $\mathbb{E}\left[P_m^R\right]$:

$$\underbrace{c_1 I_{1,L1} + c_2 I_{2,L1}}_{\bar{P}_{L1}} \leq \mathbb{E}\left[P_m^R\right] \leq \underbrace{c_1 I_{1,U1} + c_2 I_{2,U1}}_{\bar{P}_{U1}}. \tag{3.17}$$

Thus,

$$\begin{aligned} \bar{P}_{L1} = & c_1 I_{1,L1} + c_2 I_{2,L1} \\ = & c_1 \frac{1}{(M-1)}\left[1-\left(1-\rho_{th}^2\right)^{M-1}\right] \\ & + c_2 \frac{1}{(M-1)\left(1+\sqrt{\gamma_1\gamma_2}\rho_{th}^2\right)}\left[1-\left(1-\rho_{th}^2\right)^{M-1}\right] \\ = & \frac{1-(1-\rho_{th}^2)^{M-1}}{M-1}\left(c_1 + \frac{c_2}{1+\sqrt{\gamma_1\gamma_2}\rho_{th}^2}\right) = \mathcal{O}\left(\frac{1}{M}\right) \end{aligned} \tag{3.18}$$

and

$$\begin{aligned}\bar{P}_{U1} =& c_1 I_{1,U1} + c_2 I_{2,U1} \\ =& c_1 \frac{1}{(M-1)\left(1-\sqrt{\gamma_1\gamma_2}\rho_{th}^2\right)}\left[1-\left(1-\rho_{th}^2\right)^{M-1}\right] \\ &+ c_2 \frac{1}{(M-1)}\left[1-\left(1-\rho_{th}^2\right)^{M-1}\right] \\ =& \frac{1-\left(1-\rho_{th}^2\right)^{M-1}}{M-1}\left(\frac{c_1}{1-\sqrt{\gamma_1\gamma_2}\rho_{th}^2}+c_2\right) = \mathcal{O}\left(\frac{1}{M}\right).\end{aligned} \tag{3.19}$$

When $c_2 < 0$, we have

$$\underbrace{c_1 I_{1,L1} + c_2 I_{2,U1}}_{\bar{P}_{L2}} \leq \mathbb{E}\left[P_m^R\right] \leq \underbrace{c_1 I_{1,U1} + c_2 I_{2,L1}}_{\bar{P}_{U2}}. \tag{3.20}$$

Thus,

$$\begin{aligned}\bar{P}_{L2} =& c_1 I_{1,L1} + c_2 I_{2,U1} \\ =& c_1 \frac{1}{(M-1)}\left[1-\left(1-\rho_{th}^2\right)^{M-1}\right] \\ &+ c_2 \frac{1}{(M-1)\left(1+\sqrt{\gamma_1\gamma_2}\rho_{th}^2\right)}\left[1-\left(1-\rho_{th}^2\right)^{M-1}\right] \\ =& \frac{1-\left(1-\rho_{th}^2\right)^{M-1}}{M-1}\left(c_1+\frac{c_2}{1+\sqrt{\gamma_1\gamma_2}\rho_{th}^2}\right) = \mathcal{O}\left(\frac{1}{M}\right)\end{aligned} \tag{3.21}$$

and

$$\begin{aligned}\bar{P}_{U2} =& c_1 I_{1,U1} + c_2 I_{2,L1} \\ =& c_1 \frac{1}{(M-1)\left(1-\sqrt{\gamma_1\gamma_2}\rho_{th}^2\right)}\left[1-\left(1-\rho_{th}^2\right)^{M-1}\right] \\ &+ c_2 \frac{1}{(M-1)}\left[1-\left(1-\rho_{th}^2\right)^{M-1}\right] \\ =& \frac{1-\left(1-\rho_{th}^2\right)^{M-1}}{M-1}\left(\frac{c_1}{1-\sqrt{\gamma_1\gamma_2}\rho_{th}^2}+c_2\right) = \mathcal{O}\left(\frac{1}{M}\right).\end{aligned} \tag{3.22}$$

□

## 3.3 Hybrid scheme of NOMA and multi-user beamforming

According to the discussion in Section 3.1, the multi-user beamforming scheme works well with small $\rho$ and cannot guarantee any SINR requirement when $\rho$ is

large. On the other hand, the previous discussion in Chapter 2 tells us that the NOMA scheme works well with large $\rho$ and needs very large power to guarantee SINR requirement when $\rho$ is small. The two schemes are complementary. Therefore, an alignment-based hybrid scheme of NOMA and multi-user beamforming can be used to cover all scenarios with improved performance. In the alignment-based hybrid scheme, when $\rho$ is smaller than the threshold, multi-user beamforming is used while NOMA transmission is conducted when $\rho$ is larger than the threshold.

A hybrid scheme of NOMA and multi-user beamforming based on the threshold of $\rho$ was proposed in [38] and its superiority compared to the multi-user beamforming scheme in terms of spectral efficiency was shown by computer simulations. In this section, we propose several threshold selections and give theoretical analysis on the average required transmit power of the hybrid scheme.

**Algorithm 1:** The optimal hybrid scheme.

```
For each channel realization, calculate the alignment of channel
  directions ρ, instantaneous required transmit power for NOMA
  scheme P_n^R and instantaneous required transmit power for multi-user
  beamforming scheme P_m^R;
if ρ^2 ≥ 1/√(γ_H γ_T) then
    Conduct NOMA transmission;
else
    if P_m^R > P_n^R then
        Conduct NOMA transmission;
    else
        Conduct multi-user beamforming transmission;
    end
end
```

To begin with, we propose the optimal hybrid scheme to minimize the required transmit power in Algorithm 1. The scheme is based on our analytical result in (2.26) and (3.10) on the instantaneous transmit power. In this optimal hybrid scheme, the BS conducts NOMA transmission when $\rho^2 \geq 1/\sqrt{\gamma_H \gamma_T}$ as multi-user beamforming cannot guarantee the SINR levels under this circumstance; and when $\rho^2 < 1/\sqrt{\gamma_H \gamma_T}$, the scheme with a lower required transmit

power is chosen to reduce the power consumption.

The optimal hybrid scheme compares $P_m^R$ and $P_n^R$, which is not exactly a threshold based scheme. It serves as a benchmark. The hybrid scheme based on the alignment threshold is given by Algorithm 2. For this alignment-based hybrid scheme, based on the previous results on the alignment-based schemes, the average required transmit power for the hybrid scheme to guarantee the SINR levels for both user is given by

$$\bar{P}_H^R = \mathbb{E}\left[P_m^R\right] + \left(1-\rho_{th}^2\right)^{M-1}\mathbb{E}\left[P_{n,I}^R\right] \approx \mathbb{E}\left[P_m^R\right] + \left(1-\rho_{th}^2\right)^{M-1}\tilde{P}_{lo} \quad (3.23)$$

with Criterion-I and

$$\bar{P}_H^R = \mathbb{E}\left[P_m^R\right] + \left(1-\rho_{th}^2\right)^{M-1}\mathbb{E}\left[P_{n,II}^R\right] \quad (3.24)$$

with Criterion-II.

**Algorithm 2:** The alignment-based hybrid scheme.

```
For each channel realization, calculate the channel correlation
  coefficient ρ;
if ρ > ρ_th then
    Conduct NOMA transmission;
else
    Conduct multi-user beamforming transmission;
end
```

Recall the instantaneous required transmit power for NOMA scheme with Criterion-II:

$$P_{n,II}^R = \frac{\gamma_H(1+\gamma_T)}{\beta_{H_{II}}\|\boldsymbol{h}_{H_{II}}\|^2} + \frac{\gamma_T}{\beta_{T_{II}}\|\boldsymbol{h}_{T_{II}}\|^2\rho^2}. \quad (3.25)$$

As the instantaneous required transmit power is dominated by the second term, by ignoring the first term we can obtain:

$$\tilde{P}_{n,II}^R = \frac{\gamma_T}{\beta_{T_{II}}\|\boldsymbol{h}_{T_{II}}\|^2\rho^2}. \quad (3.26)$$

For the multi-user beamforming scheme, since the notation of user indices

do not affect the value of $P_m^R$, (3.10) can be rewritten as

$$\begin{aligned} P_m^R =& \frac{\gamma_H + \gamma_H\gamma_T\rho^2}{1-\gamma_H\gamma_T\rho^4}\frac{1}{\beta_{H_{II}}\|\boldsymbol{h}_{H_{II}}\|^2} + \frac{\gamma_T + \gamma_H\gamma_T\rho^2}{1-\gamma_H\gamma_T\rho^4}\frac{1}{\beta_{T_{II}}\|\boldsymbol{h}_{T_{II}}\|^2} \\ =& \frac{\gamma_H}{2}\frac{1}{\beta_{H_{II}}\|\boldsymbol{h}_{H_{II}}\|^2}\left[\frac{1-\sqrt{\frac{\gamma_T}{\gamma_H}}}{1+\sqrt{\gamma_H\gamma_T\rho^2}} + \frac{1+\sqrt{\frac{\gamma_T}{\gamma_H}}}{1-\sqrt{\gamma_H\gamma_T\rho^2}}\right] + \\ & \frac{\gamma_T}{2}\frac{1}{\beta_{T_{II}}\|\boldsymbol{h}_{T_{II}}\|^2}\left[\frac{1-\sqrt{\frac{\gamma_H}{\gamma_T}}}{1+\sqrt{\gamma_H\gamma_T\rho^2}} + \frac{1+\sqrt{\frac{\gamma_H}{\gamma_T}}}{1-\sqrt{\gamma_H\gamma_T\rho^2}}\right]. \end{aligned} \tag{3.27}$$

By ignoring the smaller terms in both the square brackets in (3.27), we can obtain:

$$\tilde{P}_m^R = \frac{1}{2\left(1-\sqrt{\gamma_H\gamma_T\rho^2}\right)}\left[\frac{\gamma_H\left(1+\sqrt{\frac{\gamma_T}{\gamma_H}}\right)}{\beta_{H_{II}}\|\boldsymbol{h}_{H_{II}}\|^2} + \frac{\gamma_T\left(1+\sqrt{\frac{\gamma_H}{\gamma_T}}\right)}{\beta_{T_{II}}\|\boldsymbol{h}_{T_{II}}\|^2}\right]. \tag{3.28}$$

Then, by comparing $\tilde{P}_R^N$ and $\tilde{P}_R^M$ we can obtain:

$$\tilde{P}_R^N \lessgtr \tilde{P}_R^M \iff$$

$$\frac{\gamma_T}{\beta_{T_{II}}\|\boldsymbol{h}_{T_{II}}\|^2\rho^2} \lessgtr \frac{1}{2\left(1-\sqrt{\gamma_H\gamma_T\rho^2}\right)}\left[\frac{\gamma_H\left(1+\sqrt{\frac{\gamma_T}{\gamma_H}}\right)}{\beta_{H_{II}}\|\boldsymbol{h}_{H_{II}}\|^2} + \frac{\gamma_T\left(1+\sqrt{\frac{\gamma_H}{\gamma_T}}\right)}{\beta_{T_{II}}\|\boldsymbol{h}_{T_{II}}\|^2}\right]$$

$$\iff \rho^2 \gtrless \rho_{th,1}^2 \triangleq \frac{1}{\sqrt{\gamma_H\gamma_T} + \frac{\sqrt{\gamma_H}+\sqrt{\gamma_T}}{2}\left(\frac{1}{\sqrt{\gamma_T}} + \frac{\sqrt{\gamma_H}}{\gamma_T}\frac{\beta_{T_{II}}\|\boldsymbol{h}_{T_{II}}\|^2}{\beta_{H_{II}}\|\boldsymbol{h}_{H_{II}}\|^2}\right)}.$$

This leads to our first threshold design.

The value of $\rho_{th,1}^2$ depends on $\|\boldsymbol{h}_{H_{II}}\|^2$ and $\|\boldsymbol{h}_{T_{II}}\|^2$, thus the BS needs to calculate $\rho_{th,1}^2$ for every coherent interval, which is similar to the optimal hybrid scheme though less computations. Therefore, we propose the following fixed threshold design. By replacing $\frac{\beta_{T_{II}}\|\boldsymbol{h}_{T_{II}}\|^2}{\beta_{H_{II}}\|\boldsymbol{h}_{H_{II}}\|^2}$ with $\mathbb{E}\left[\frac{\beta_{T_{II}}\|\boldsymbol{h}_{T_{II}}\|^2}{\beta_{H_{II}}\|\boldsymbol{h}_{H_{II}}\|^2}\right]$, we can obtain

$$\rho_{th,2}^2 \triangleq \frac{1}{\sqrt{\gamma_H\gamma_T} + \frac{1}{2}\left(1+\sqrt{\frac{\gamma_H}{\gamma_T}}\right)\left(1+\sqrt{\frac{\gamma_H}{\gamma_T}}\mathbb{E}\left[\frac{\beta_{T_{II}}\|\boldsymbol{h}_{T_{II}}\|^2}{\beta_{H_{II}}\|\boldsymbol{h}_{H_{II}}\|^2}\right]\right)}. \tag{3.29}$$

Finally, simply consider the condition in (3.11) we can obtain a benchmark threshold design.

$$\rho_{th,3}^2 \triangleq \frac{1}{\sqrt{\gamma_H\gamma_T}}. \tag{3.30}$$

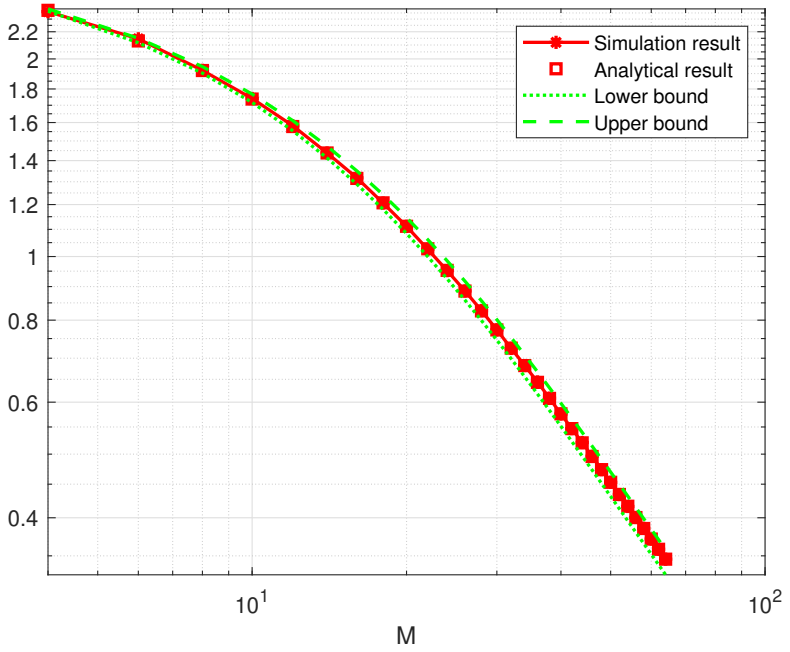


Figure 3.1: Average required transmit power for the alignment-based multi-user beamforming scheme versus $M$.

## 3.4 Simulation results

In this section, computer simulations are implemented to validate the analytical result on the power consumption of multi-user beamforming, as well as the performance of the hybrid schemes on power saving.

### 3.4.1 Power consumption of the alignment-based multi-user beamforming scheme

Fig. 3.1 shows the average required transmit power for the alignment-based multi-user beamforming scheme and its bounds versus $M$ where $\rho_{th} = 0.3$, $\gamma_1 = 10\text{dB}$, $\gamma_2 = 0\text{dB}$, $\beta_1 = 0\text{dB}$ and $\beta_2 = -10\text{dB}$. The analytical result is obtained by (3.12) while the lower and upper bounds are obtained by (3.16). We can see that the analytical result matches perfectly with the simulation result and the bounds are very close to the simulation result, which demonstrate the correctness of Theorem 3.1 and Corollary 3.1. Another observation is that the average required transmit power decreases with $1/M$ when $M$ is large, which validates the asymptotic behavior of the average required transmit power given

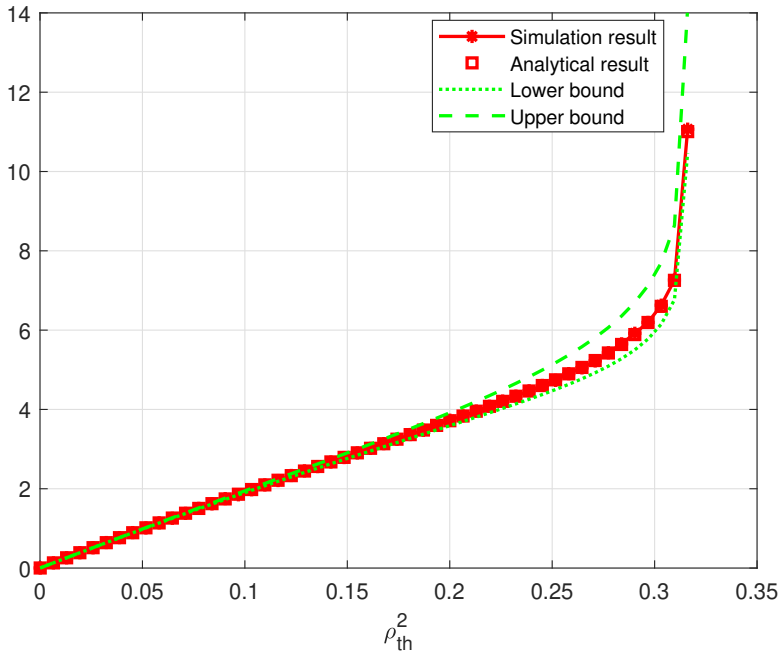


Figure 3.2: Average required transmit power for the alignment-based multi-user beamforming scheme versus $\rho_{th}^2$.

by Corollary 3.2.

In Fig. 3.2, the average required transmit power for the alignment-based multi-user beamforming scheme and its bounds are shown as functions of $\rho_{th}^2$ where $M = 8$, $\gamma_1 = 10$dB, $\gamma_2 = 0$dB, $\beta_1 = 0$dB and $\beta_2 = -10$dB. The correctness of Theorem 3.1 is demonstrated again. And we can find that the bounds are tight even when $\rho_{th}^2$ is close to $1/\sqrt{\gamma_1\gamma_2}$.

### 3.4.2 Power consumption of the hybrid scheme

Fig. 3.3 shows the average required transmit power for the alignment-based hybrid scheme versus $\rho_{th}^2$ where $M = 8$, $\gamma_1 = 10$dB, $\gamma_2 = 2$dB, $\beta_1 = 0$dB and $\beta_2 = -10$dB. We also plot the values of $\rho_{th,2}^2$ and $\rho_{th,3}^2$ by the vertical dashed lines. It can be seen that the average required transmit power decreases firstly and then increases with $\rho_{th}^2$, indicating that the hybrid scheme can save power. For our threshold designs, we can find that $\rho_{th,2}$ is near optimal on power saving.

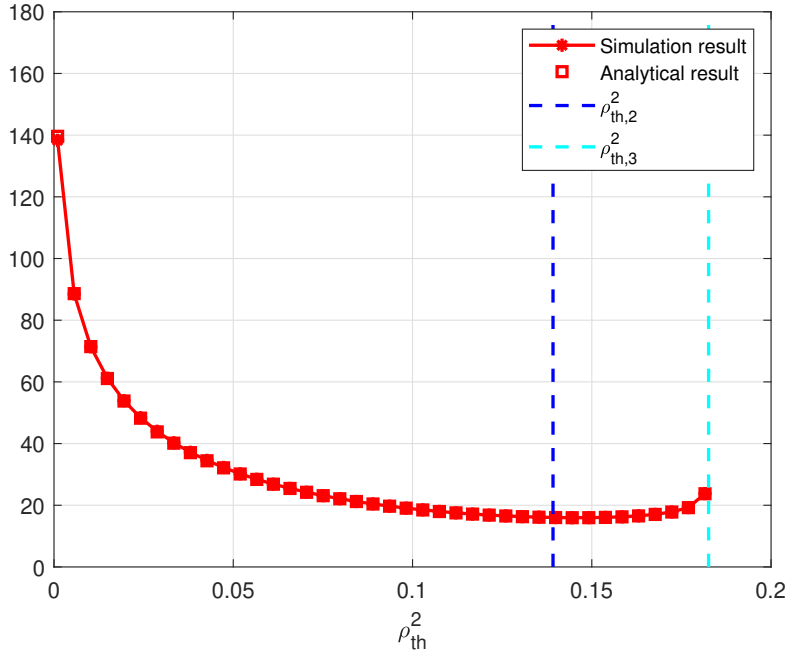


Figure 3.3: Average required transmit power for the alignment-based hybrid scheme versus $\rho^2_{th}$.

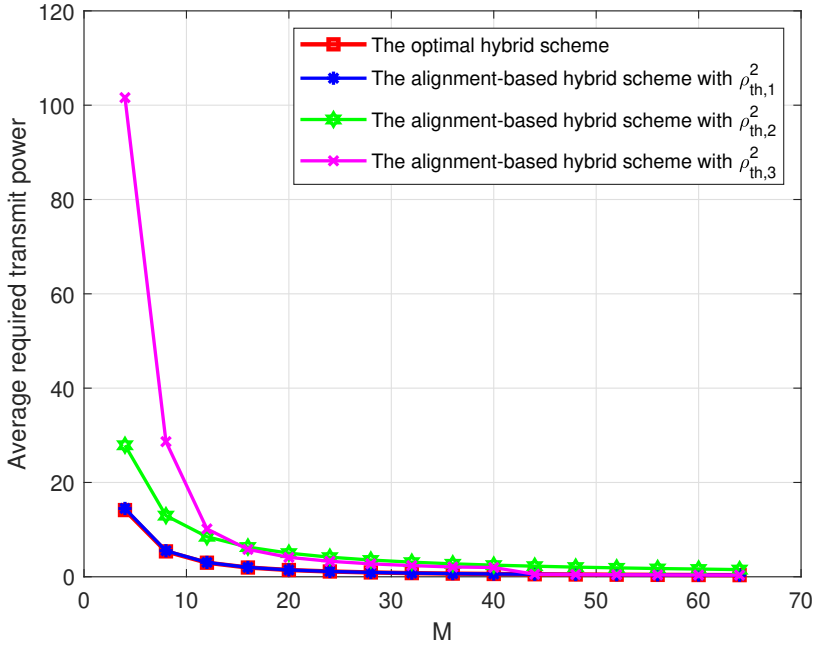


Figure 3.4: Average required transmit power for the optimal and alignment-based hybrid schemes versus $M$.

Fig. 3.4 shows the average required transmit power versus $M$ for the optimal hybrid scheme and alignment-based hybrid scheme with different threshold designs. We can find that the optimal hybrid scheme provide the least average required transmit power and the alignment-based hybrid scheme with $\rho_{th,1}^2$ performs nearly the same as the optimal hybrid scheme. Another observation is the average required transmit power for the alignment-based hybrid scheme with $\rho_{th,3}^2$ is the biggest when $M$ is small and decrease rapidly with $M$. When $M$ is large enough, the alignment-based hybrid scheme with $\rho_{th,3}^2$ consumes less power than the alignment-based hybrid scheme with $\rho_{th,2}^2$ and performs closely to the optimal hybrid scheme.

## 3.5 Conclusion

In this chapter the power consumption of multi-user beamforming and the hybrid of NOMA and multi-user beamforming is investigated. By deriving the instantaneous required transmit power for multi-user beamforming scheme to guarantee the SINR levels for both users, we find that multi-user beamforming can play as a complement of the NOMA scheme since it works well with small alignment of channel directions. Thus a hybrid of the two schemes is desirable and we proposed a hybrid scheme with several alignment threshold designs. Simulation results illustrate the theoretical result on multi-user beamforming and the hybrid scheme.

# Chapter 4

# User clustering design for multi-user NOMA systems

The work on NOMA systems in Chapter 2 is based on the two-user scenario. For the multi-user scenario, two directions can be considered. One direction is to group users into multiple two-user clusters with NOMA inside each cluster, i.e., the two users in the same cluster share the common time-frequency resource block and perform SIC to eliminate the intra-cluster interference, which is called multi-cluster NOMA system[1]. The other is to consider all users in one cluster and such a system is called single-cluster NOMA system.

In this chapter, we take the first direction to work on the user-pairing design aiming at power saving in multi-cluster multi-antenna NOMA systems. The user-pairing problem is formulated to minimize the total required transmit power under given SINR constraints for all users. Two algorithms, one optimal and the other suboptimal, are proposed to solve the user-clustering problem. The optimal algorithm clusters users by solving a maximum cardinality minimum weight matching problem over a graph constructed by the system model. The sub-optimal algorithm has two stages where the first stage chooses the head or strong user in each cluster according to its effective channel gain and the second stage pairs the remaining users by the Hungarian algorithm. Significant power saving is achieved by the proposed user-cluster solutions.

[1]In general, the multi-cluster NOMA system is not restricted to the two-user cluster case, each cluster can have multiple users and the two-user cluster case is studied in this chapter.

The remainder of this chapter is organized as follow. The user-clustering problem is formulated and solved by two algorithms in Section 4.1. Section 4.2 contains the transmit power analysis of the solutions found by two algorithms. Section 4.3 shows the numerical results including both the total required transmit power and run-time of the algorithms.

## 4.1 The clustering problem and solutions for multi-user multi-antenna NOMA

Consider the downlink transmission from an $M$-antenna BS to $2K$ single-antenna users and the channel model is given by (1.3) and (1.4). In our multi-cluster multi-antenna NOMA scheme, all users are firstly grouped into $K$ clusters with two users per cluster, one head user and one tail user, and the intra-cluster interference can be avoided by SIC within each cluster. The inter-cluster interference can be eliminated either by adopting appropriate beamforming schemes such as the zero-forcing beamforming adopted in [38] or allocating orthogonal time-frequency resource block to different clusters. In this chapter, we consider the latter method. Therefore, orthogonal-resource blocks are assigned to the $K$ clusters and in each cluster the BS serves two users with common time-frequency resource block as well as common beamformer.

Denote the required transmit power for the $i$th cluster as $P_{C,i}$. Since the inter-cluster interference is avoided by orthogonal resource allocation, by adopting MF beamforming with respect to the head users and decoding order determined by Criterion-II, we have:

$$P_{C,i} \triangleq \frac{\gamma_H(1+\gamma_T)}{\beta_{H_i}\|\boldsymbol{h}_{H_i}\|^2} + \frac{\gamma_T}{\beta_{T_i}\|\boldsymbol{h}_{T_i}\|^2\rho_i^2}, \tag{4.1}$$

where $H_i$ and $T_i$ are the indices of the head and tail users in the $i$th cluster; $\rho_i$ defined by

$$\rho_i \triangleq \frac{\left|\boldsymbol{h}_{H_i}^H \boldsymbol{h}_{T_i}\right|}{\|\boldsymbol{h}_{H_i}\|\|\boldsymbol{h}_{T_i}\|}.$$

is the alignment of channel directions in the $i$th cluster.

Define

$$P_{C,i}^{(1)} \triangleq \frac{\gamma_H(1+\gamma_T)}{\beta_{H_i}\|\boldsymbol{h}_{H_i}\|^2}, \tag{4.2}$$

and

$$P_{C,i}^{(2)} \triangleq \frac{\gamma_T}{\beta_{T_i}\|\boldsymbol{h}_{T_i}\|^2\rho_i^2}. \tag{4.3}$$

From (4.1), the total BS transmit power for the user SINR guarantee is thus:

$$P_{total} = \sum_{i=1}^{K} P_{C,i} = \sum_{i=1}^{K} \left(P_{C,i}^{(1)} + P_{C,i}^{(2)}\right). \tag{4.4}$$

Denote the set of user indices as $\mathcal{U} = \{1, 2, \cdots, 2K\}$. It can be seen that the total BS transmit power depends on the user clustering, represented by a permutation of $\mathcal{U}$: $(H_1, H_2, \cdots, H_K, T_1, T_2, \cdots, T_K)$. The user clustering problem can thus be formulated as follows:

$$\text{P1}: \quad \min_{(H_1,\cdots,H_K,T_1,\cdots,T_K)\in S_{2K}} \sum_{i=1}^{K} \left(P_{C,i}^{(1)} + P_{C,i}^{(2)}\right),$$

where $S_{2K}$ is the set of all permutations of $\mathcal{U}$.

The user-cluster problem can be transformed into the matching problem in graph theory. Given a undirected weighted graph $G = (V, E, w)$ where $V$ is the set of vertices, $E$ is the set of edges, $w$ is the set of weights of the edges. A *matching* is a subset of edges $E' \subseteq E$ such that each node in $V$ has at most one incident edge in $E'$. The goal of the maximum cardinality minimum weight matching problem is to find a matching $E'$ with the maximum cardinality $|E'|$ and the minimum weight $w(E')$. Therefore, we can solve P1 by solving the maximum cardinality minimum weight matching problem over graph $G_c = (V_c, E_c, w_c)$ where each vertex in $V_c$ represents a user, any two users are connected by an edge in $E_c$, and the weight on each edge is calculated by (4.1). An optimal algorithm for this problem is provided in [48] with complexity $\mathcal{O}(K^3 \log K)$. This algorithm is referred to as the optimal algorithm.

The optimal algorithm is very complicated and performance analysis is high challenging if not impossible. As will be shown in Section 4.3, the computational complexity of the optimal algorithm is also high. In what follows,

we propose a suboptimal algorithm for the user clustering. By noticing that $P_{C,i}^{(1)}$ defined in (4.2) only depend on the head users $H_i$'s, the total transmit power in (4.4) can be rewritten as

$$P_{total} = \sum_{i=1}^{K} P_{C,i}^{(1)} + \sum_{i=1}^{K} P_{C,i}^{(2)}.$$

This inspires the following transformation to simplify the clustering problem:

$$\begin{aligned} \text{P2}: \min_{(T_1,\cdots,T_K)\in S_{2K}} & \left[ \left( \min_{(H_1,\cdots,H_K)} \sum_{i=1}^{K} P_{C,i}^{(1)} \right) + \sum_{i=1}^{K} P_{C,i}^{(2)} \right] \\ \text{s.t.} \quad & (H_1,\cdots,H_K,T_1,\cdots,T_K) \in S_{2K}. \end{aligned}$$

It should be noted that the transformation results in sub-optimality of the solution, which is leveraged for the complexity consideration.

The problem P2 naturally leads to a two-layer clustering design, where in the first layer, head users are chosen to minimize $\sum_{i=1}^{K} P_{C,i}^{(1)}$ (i.e., solving the inner sub-problem) and in the second layer, tail users are chosen to minimize the total transmit power. This two-layer algorithm is referred to as the sub-optimal algorithm. For Layer 1, the optimal head users $(H_1,\cdots,H_K)$ can be found by ordering $\beta_i\|\boldsymbol{h}_i\|^2$'s in the descending order and pick the first half users as the cluster heads. That is, by ordering the users such as

$$\beta_{\sigma_1}\|\boldsymbol{h}_{\sigma_1}\|^2 \geq \beta_{\sigma_2}\|\boldsymbol{h}_{\sigma_2}\|^2 \geq \cdots \geq \beta_{\sigma_K}\|\boldsymbol{h}_{\sigma_{2K}}\|^2,$$

we have $H_i = \sigma_i$ for $i = 1,\cdots,K$ as the indices of head users and the remaining later half are the tail users.

For Layer 2, the optimization problem is equivalent to

$$\begin{aligned} \text{P3}: \min_{(T_1,\cdots,T_K)} & \left( \sum_{i=1}^{K} P_{C,i}^{(2)} \right) \\ \text{s.t.} \ & (T_1,\cdots,T_K) \text{ is a permutation of } (\sigma_{K+1},\cdots\sigma_{2K}), \end{aligned}$$

which is an *assignment problem*, i.e., assign the remaining users to the head users to minimize the sum of the cost. Therefore, we can solve P3 by the Hungarian algorithm [49]–[51] with complexity $\mathcal{O}(K^3)$ and the Hungarian algorithm is described in Algorithm 3. The cost matrix, which is also the input

**Algorithm 3:** Hungarian algorithm for P3.

**Input:** Cost matrix $\mathbf{\Lambda}$
**Output:** Assignment set $\Omega$

```
m = 0;
Find and subtract the minimum elements from all elements in each row;
Find and subtract the minimum elements from all elements in each column;
while m = 0 do
    Find the minimum number of columns and rows N to cover all the 0 elements of Λ;
    if N = K then
        m = 1;
    else
        Find the minimum uncovered element S;
        Subtract S from each uncovered row and add S to each covered column;
    end
end
Choose K 0 elements from different columns and rows. The indices of these chose elements are the assignment result.
```

of the assignment algorithms, is defined by:

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_{1,1} & \cdots & \lambda_{1,K} \\ \vdots & \ddots & \vdots \\ \lambda_{K,1} & \cdots & \lambda_{K,K} \end{pmatrix},$$

with the $(i, j)$th entry given by:

$$\lambda_{i,j} = \frac{\gamma_T}{\beta_{\sigma_{K+j}} \|\boldsymbol{h}_{\sigma_{K+j}}\|^2 \rho^2_{H_i,\sigma_{K+j}}}.$$

It is worthy reminding here that $H_i = \sigma_i$ is the index of the head user of Cluster $i$; $\sigma_{K+j}$ is the index of the $j$th unassigned user; $\rho^2_{H_i,\sigma_{K+j}}$ is the alignment of channel directions of user $H_i$ and $\sigma_{K+j}$; $\lambda_{i,j}$ represents the cost of assigning $\sigma_{K+j}$ as the tail user of Cluster $i$.

## 4.2 Performance analysis

The following theorem is proved for the average required transmit power for the NOMA scheme with our proposed sub-optimal user clustering algorithm.

Naturally, it also works as a lower bound for the performance of the optimal scheme.

**Theorem 4.1** *When $M \geq 2$, and $K \geq 2$, the average required transmit power for the NOMA scheme with our proposed sub-optimal user clustering algorithm to guarantee the SINR levels for all users is bounded.*

*Proof:* In Layer 1 of the sub-optimal algorithm, we minimize

$$\sum_{i=1}^{K} P_{C,i}^{(1)} = \sum_{i=1}^{K} \frac{\gamma_H(1+\gamma_T)}{\beta_{H_i}\|\boldsymbol{h}_{H_i}\|^2}$$

by ordering the users' large scale fading coefficients in descending order and pick the first $K$ users as the head users. Therefore, the mean value of $\sum_{i=1}^{K} P_{C,i}^{(1)}$ can be calculated by:

$$\mathbb{E}\left[\sum_{i=1}^{K} P_{C,i}^{(1)}\right] = \gamma_H\left(1+\gamma_T\right)\sum_{i=1}^{K}\mathbb{E}\left[\frac{1}{\beta_{H_i}\|\boldsymbol{h}_{H_i}\|^2}\right] < \frac{\gamma_H\left(1+\gamma_T\right)}{M-1}\sum_{i=1}^{K}\frac{1}{\beta_{H_i}} < \infty.$$

In Layer 2 of the sub-optimal algorithm, we minimize

$$\sum_{i=1}^{K} P_{C,i}^{(2)} = \sum_{i=1}^{K} \frac{\gamma_T}{\beta_{T_i}\|\boldsymbol{h}_{T_i}\|^2\rho_i^2}$$

by the Hungarian algorithm. Since the solution of the Hungarian algorithm does not have an explicit expression, it is difficult to analyze the performance directly. Instead, we discuss the worst case which can be an upper bound on the transmit power of the solution found by the Hungarian algorithm.

In the worst case, the tail users always collide with each other when choosing their head users, i.e., all the tail users have the same 1st, 2nd, ..., $K$-th choices of head users. Since the Hungarian algorithm can find the optimal solution with the minimum cost based on the cost matrix $\Lambda$, any specific assignment based on the following cost matrix

$$\boldsymbol{\Delta} = \begin{pmatrix} \frac{1}{\rho_{H_1,\sigma_{K+1}}^2} & \cdots & \frac{1}{\rho_{H_1,\sigma_{K+K}}^2} \\ \vdots & \ddots & \vdots \\ \frac{1}{\rho_{H_K,\sigma_{K+1}}^2} & \cdots & \frac{1}{\rho_{H_K,\sigma_{K+K}}^2} \end{pmatrix} \tag{4.5}$$

must has a larger cost than the Hungarian algorithm. In the following, we analyze the performance of an assignment based on $\Delta$.

Recall that in the worst case scenario, all tail users have the same worst head user. Denote the worst head user as $H_w$, then the following equation holds for all $j = \sigma_{K+1}, \sigma_{K+2} \cdots, \sigma_{2K}$

$$\rho^2_{H_w,j} = \min\left(\rho^2_{H_1,j}, \rho^2_{H_2,j}, \cdots, \rho^2_{H_K,j}\right).$$

One assignment is to match the best tail user (denote as user $T_b$) to $H_w$ so that

$$\rho^2_{H_w,T_b} = \max\left(\rho^2_{H_w,\sigma_{K+1}}, \rho^2_{H_w,\sigma_{K+2}}, \cdots, \rho^2_{H_w,\sigma_{2K}}\right).$$

For the other tail users, since the worst head users has been assigned, the average transmit power of any user pair should be smaller than $\mathbb{E}\left[\frac{1}{\rho^2_{H_w,T_b}}\right]$.

Since $\rho^2_{H_w,j}$ is the smallest value among $K$ independent samples following Beta$(1, M-1)$. The PDF of $\rho^2_{H_w,j}$ is given by:

$$\begin{aligned} f_{\rho^2_{H_w,j}}(x) =& K\left[1 - F_{\rho^2}(x)\right]^{K-1} f_{\rho^2}(x) \\ =& K(M-1)\left(1-x\right)^{K(M-1)-1}, \end{aligned} \tag{4.6}$$

which is the same as the PDF of Beta$(1, K(M-1))$. Since $\rho^2_{H_w,T_b}$ is the largest value among $K$ independent samples following Beta$(1, K(M-1))$ and its PDF can be calculated by:

$$\begin{aligned} f_{\rho^2_{H_w,T_b}}(x) =& K\left[F_{\rho^2_{H_w,j}}(x)\right]^{K-1} f_{\rho^2_{H_w,j}}(x) \\ =& K^2(M-1)\left[1-(1-x)^{K(M-1)}\right]^{K-1}(1-x)^{K(M-1)-1}. \end{aligned} \tag{4.7}$$

Then, an upper bound of the cost is given by:

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{K} \frac{\gamma_T}{\beta_{T_i}\|\boldsymbol{h}_{T_i}\|^2\rho_i^2}\right] \leq& \mathbb{E}\left[\sum_{i=1}^{K} \frac{\gamma_T}{\beta_{T_i}\|\boldsymbol{h}_{T_i}\|^2}\right] \cdot \mathbb{E}\left[\frac{1}{\rho^2_{m,n}}\right] \\ =& \gamma_T K^2(M-1)B \cdot \mathbb{E}\left[\sum_{i=1}^{K} \frac{1}{\beta_{T_i}\|\boldsymbol{h}_{T_i}\|^2}\right] \end{aligned} \tag{4.8}$$

and $B$ can be calculated by

$$\begin{aligned}
B =& \int_0^1 \frac{\left[1-(1-x)^{K(M-1)}\right]^{K-1}}{x}(1-x)^{K(M-1)-1}\,\mathrm{d}x \\
\overset{(e)}{=}& \int_0^1 \frac{\left[1-y^{K(M-1)}\right]^{K-1}}{1-y}y^{K(M-1)-1}\,\mathrm{d}y \\
=& \int_0^1 \left(\sum_{i=0}^{K(M-1)-1} y^i\right)\left(1-y^{K(M-1)}\right)^{K-2} y^{K(M-1)-1}\,\mathrm{d}y \\
\leq& \left(\sum_{i=0}^{K(M-1)-1} 1^i\right)\int_0^1 \left(1-y^{K(M-1)}\right)^{K-2} y^{K(M-1)-1}\,\mathrm{d}y \\
\leq& K(M-1)\int_0^1 y^{K(M-1)-1}\,\mathrm{d}y \\
=& \left. y^{K(M-1)-1}\right|_0^1 = 1,
\end{aligned} \tag{4.9}$$

where $(e)$ is obtained by the change of variable $y = 1 - x$.

Therefore, we have

$$\mathbb{E}\left[\sum_{i=1}^{K} P_{C,i}^{(2)}\right] = \mathbb{E}\left[\sum_{i=1}^{K} \frac{\gamma_T}{\beta_{T_i}\|\boldsymbol{h}_{T_i}\|^2\rho_i^2}\right] < \infty,$$

thus

$$\mathbb{E}\left[\sum_{i=1}^{K} P_{C,i}^{(1)}\right] + \mathbb{E}\left[\sum_{i=1}^{K} P_{C,i}^{(2)}\right] < \infty,$$

which concludes the proof. □

## 4.3 Simulation results

In this section, the simulation results are given to show the performance of the user clustering algorithms. We set $\gamma_H = 10\text{dB}$ and $\gamma_T = 0\text{dB}$ considering the difference of channel conditions. The large scale coefficients are randomly and independently generated following the uniform distribution on $(-15, 0)$, i.e., $\beta_k(\text{dB}) \sim \text{U}(-15, 0)$. Four user clustering schemes are studied: the optimal algorithm, the two-layer sub-optimal algorithm, the random clustering where both cluster heads and cluster tails are selected randomly, and a semi-random clustering where the head users are chosen following Layer 1 of the two-layer sub-optimal algorithm while the tail users are assigned randomly.

Fig. 4.1 shows the average required transmit power for these user clustering schemes for a network with 8 users, i.e., K=4. It can be seen that the sub-optimal algorithm performs closely to the optimal algorithm and has significant power saving compared to the random clustering as well as the semi-random clustering. Further, the average total transmit power of the sub-optimal and optimal algorithms decreases with $M$, while this cannot be said for the random and semi-random ones. The average total power of the random and semi-random ones suffers huge fluctuations. This can be explained as follows. As shown in Corollary 2.1 and Corollary 2.5, the average transmit power for a multi-antenna NOMA system with no restriction on the channel correlation of users in the same cluster is unbounded. The random and semi-random user clustering schemes cannot address this issue, thus their average total power can have big difference from one channel realization to another. On the other hand, the sub-optimal and optimal algorithms avoids the pair of users with near-orthogonal channel vectors.

Fig. 4.2 compares the run-time of the sub-optimal and the optimal algorithm with respect to $K$. A significant difference between the rum-time of the optimal and sub-optimal algorithms can be observed and the superiority of the sub-optimal algorithm is demonstrated in terms of computation complexity. The optimal algorithm is not efficient with large $K$ since it may bring high latency in communications.

## 4.4 Conclusion

This chapter investigated the power consumption and user-clustering of a multi-cluster multi-antenna NOMA system. By transforming the clustering problem into the matching problem in graph theory, we found the optimal clustering scheme with minimum total transmit power by solving the maximum cardinality minimum weight matching problem. By decomposing the transmit power formula into two parts, we developed a two-layer sub-optimal user clustering algorithm to cluster users into multiple two-user clusters targeting at minimizing the total transmit power. In the first layer, the cluster

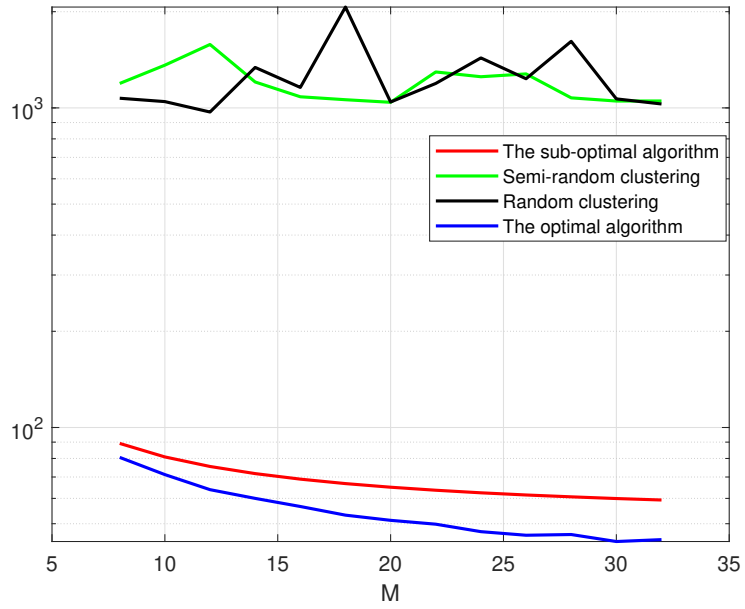


Figure 4.1: Total average transmit power versus $M$ where $K = 4$, $\gamma_H = 10$dB, $\gamma_T = 0$dB.

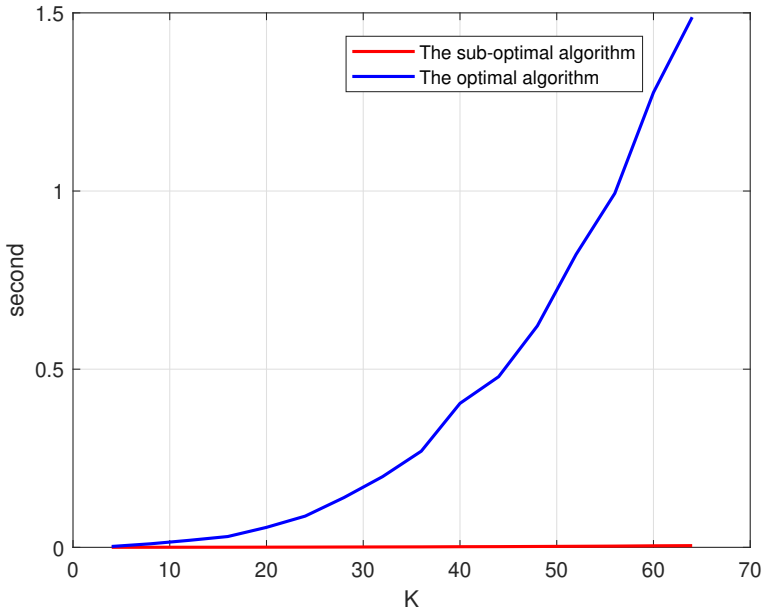


Figure 4.2: The run-time of the optimal and sub-optimal algorithms versus $K$

heads are chosen based on the strength of the channel gains; while in the second layer, Hungarian algorithm is used to pair each cluster head with a tail user to minimize the second term of the total transmit power. Theoretical result shows that both the optimal and sub-optimal algorithms can guarantee the SINR levels for all users with finite average transmit power. Numerical result demonstrate the significant improvement of our proposed clustering algorithms in multi-user multi-antenna NOMA systems.

# Chapter 5

# Conclusion and future work

In this thesis we investigate the power consumption of power domain non-orthogonal multiple access scheme with SINR guaranteed for all users in both signal cluster and multiple cluster cases.

For signal cluster case, we proposed an alignment-based NOMA transmission scheme in order to save power and derive the average required transmit power and its asymptotic behaviors with the scheme, which reveals the significance of the alignment of channel directions in NOMA system as we proved that the original NOMA without alignment threshold cannot guarantee the SINR levels for both users with finite average power. As a complement of NOMA scheme, the multi-user beamforming scheme was investigated. We derived the instantaneous and average required transmit power of multi-user beamforming scheme and demonstrated that this scheme works well with small alignment of channel directions, which motivates the hybrid scheme of NOMA and multi-user beamforming.

For multi-cluster case, two clustering algorithms are developed in order to save transmit power by transforming the clustering problem into matching problems over general and bipartite graphs in graph theory. We proved analytically that both algorithms can guarantee the SINR levels with finite average transmit power. Numerical results demonstrate the significant improvement of our proposed clustering algorithms in power saving and show the superiority of the sub-optimal algorithm in computational complexity.

In the following, we list several possible future research directions based on

the works of this thesis:

- In Chapter 2, we proposed the alignment-based scheme for clusters with two users and analyzed the average required transmit power. This work can be extended to clusters with more than two users and the alignment threshold can be set on all user pairs.

- The result of this thesis is only valid for matched filter beamforming with respect to the stronger user. The power consumption of NOMA with other beamforming scheme is worth exploring, such as the zero-forcing beamforming with respect to the stronger user [38], the matched filter beamforming based on the linear combination of the channel vectors of all users in the same cluster [42] and the beamforming scheme based on the QR decomposition of channel matrix which can create a significant difference on effective channel gains [44].

# References


[1] J. C. Maxwell, "Viii. A dynamical theory of the electromagnetic field," *Philosophical transactions of the Royal Society of London*, no. 155, pp. 459–512, 1865.

[2] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022," *Cisco white paper*, p. 9, 2016.

[3] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE 77th Veh. Technol. Conf. (VTC Spring)*, 2013, pp. 1–5.

[4] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing—a key technology towards 5G," *ETSI white paper*, vol. 11, no. 11, pp. 1–16, 2015.

[5] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 2017.

[6] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590–3605, 2016.

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

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

[9] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, 2014.

[10] T. L. Marzetta, *Fundamentals of massive MIMO*. Cambridge University Press, 2016.

[11] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 94–100, 2017.

[12] N. Alliance, "Description of network slicing concept," *NGMN 5G P*, vol. 1, no. 1, 2016.

[13] H. Zhang, N. Liu, X. Chu, K. Long, A.-H. Aghvami, and V. C. Leung, "Network slicing based 5G and future mobile networks: Mobility, resource management, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 138–145, 2017.

[14] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2429–2453, 2018.

[15] Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos, "A survey of millimeter wave communications (mmWave) for 5G: Opportunities and challenges," *Wireless Networks*, vol. 21, no. 8, pp. 2657–2676, 2015.

[16] S. Kutty and D. Sen, "Beamforming for millimeter wave communications: An inclusive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 949–973, 2015.

[17] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. K. Karagiannidis, E. Björnson, K. Yang, I. Chih-Lin, *et al.*, "Millimeter wave communications for future mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1909–1935, 2017.

[18] R. Mendez-Rial, C. Rusu, N. González-Prelcic, A. Alkhateeb, and R. W. Heath, "Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches?" *IEEE Access*, vol. 4, pp. 247–267, 2016.

[19] ITU-R, "IMT vision–framework and overall objectives of the future development of IMT for 2020 and beyond," ITU, Tech. Rep., 2015.

[20] A. Osseiran, J. F. Monserrat, and P. Marsch, *5G mobile and wireless communications technology*. Cambridge University Press, 2016.

[21] ITU-R, "Minimum requirements related to technical performance for IMT-2020 radio interface(s)," ITU, Tech. Rep., 2017.

[22] Y. Okumura, "Field strength and its variability in VHF and UHF land-mobile radio service," *Rev. Electr. Commun. Lab.*, vol. 16, pp. 825–873, 1968.

[23] W. C. Lee, *Mobile communications design fundamentals*. John Wiley & Sons, 2010, vol. 25.

[24] G. L. Stüber and G. L. Stèuber, *Principles of mobile communication*. Springer, 1996, vol. 2.

[25] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proc. IEEE 24th Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, IEEE, 2013, pp. 332–336.

[26] S. Zhang, X. Xu, L. Lu, Y. Wu, G. He, and Y. Chen, "Sparse code multiple access: An energy efficient uplink approach for 5G wireless systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, IEEE, 2014, pp. 4782–4787.

[27] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *Proc. IEEE 80th Veh. Technol. Conf. (VTC Fall)*, IEEE, 2014, pp. 1–5.

[28] S. Chen, B. Ren, Q. Gao, S. Kang, S. Sun, and K. Niu, "Pattern division multiple access—a novel nonorthogonal multiple access for fifth-generation radio networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3185–3196, 2016.

[29] X. Dai, Z. Zhang, B. Bai, S. Chen, and S. Sun, "Pattern division multiple access: A new multiple access technology for 5G," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 54–60, 2018.

[30] L. Ping, L. Liu, and W. Leung, "A simple approach to near-optimal multiuser detection: Interleave-division multiple-access," in *Proc. IEEE Wireless Commun. Networking (WCNC 2003)*, IEEE, vol. 1, 2003, pp. 391–396.

[31] L. Ping, L. Liu, K. Wu, and W. K. Leung, "Interleave division multiple-access," *IEEE Trans. Wireless Commun.*, vol. 5, no. 4, pp. 938–947, 2006.

[32] K. Kusume, G. Bauch, and W. Utschick, "IDMA vs. CDMA: Analysis and comparison of two multiple access schemes," *IEEE Trans. Wireless Commun.*, vol. 11, no. 1, pp. 78–87, 2011.

[33] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, 2016.

[34] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, 2014.

[35] Z. Chen, Z. Ding, X. Dai, and R. Zhang, "An optimization perspective of the superiority of NOMA compared to conventional OMA," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 5191–5202, 2017.

[36] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "On the sum rate of MIMO-NOMA and MIMO-OMA systems," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 534–537, 2017.

[37] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, 2017.

[38] K. Senel, H. V. Cheng, E. Björnson, and E. G. Larsson, "What role can NOMA play in massive MIMO?" *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 597–611, 2019.

[39] B. Makki, K. Chitti, A. Behravan, and M. Alouini, "A survey of NOMA: Current status and open research challenges," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 179–189, 2020.

[40] Y. Huang, C. Zhang, J. Wang, Y. Jing, L. Yang, and X. You, "Signal processing for MIMO-NOMA: Present and future challenges," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 32–38, 2018.

[41] X. Chen, F. Gong, G. Li, H. Zhang, and P. Song, "User pairing and pair scheduling in massive MIMO-NOMA systems," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 788–791, 2018.

[42] H. V. Cheng, E. Björnson, and E. G. Larsson, "Performance analysis of NOMA in training-based multiuser MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 372–385, 2018.

[43] Y. Gao, B. Xia, K. Xiao, Z. Chen, X. Li, and S. Zhang, "Theoretical analysis of the dynamic decode ordering SIC receiver for uplink NOMA systems," *IEEE Commun. Lett.*, vol. 21, no. 10, pp. 2246–2249, 2017.

[44] Z. Ding, L. Dai, and H. V. Poor, "MIMO-NOMA design for small packet transmission in the internet of things," *IEEE Access*, vol. 4, pp. 1393–1405, 2016.

[45] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in rayleigh flat fading," *IEEE Trans. Inf. Theory*, vol. 45, no. 1, pp. 139–157, 1999.

[46] Z. Chen, Z. Ding, X. Dai, and G. K. Karagiannidis, "On the application of quasi-degradation to MISO-NOMA downlink," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6174–6189, 2016.

[47] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*. Academic press, 2014.

[48] Z. Galil, "Efficient algorithms for finding maximum matching in graphs," *ACM Computing Surveys (CSUR)*, vol. 18, no. 1, pp. 23–38, 1986.

[49] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[50] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the society for industrial and applied mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[51] F. Bourgeois and J.-C. Lassalle, "An extension of the munkres algorithm for the assignment problem to rectangular matrices," *Communications of the ACM*, vol. 14, no. 12, pp. 802–804, 1971.