Resource Allocation for NOMA Wireless Systems

by

©Ming Zeng

A dissertation submitted to the School of Graduate Studies
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Faculty of Engineering and Applied Science
Memorial University of Newfoundland

May 2020

St. John's, Newfoundland

Abstract

Power-domain non-orthogonal multiple access (NOMA) has been widely recognized as a promising candidate for the next generation of wireless communication systems. By applying superposition coding at the transmitter and successive interference cancellation at the receiver, NOMA allows multiple users to access the same time-frequency resource in power domain. This way, NOMA not only increases the system's spectral and energy efficiencies, but also supports more users when compared with the conventional orthogonal multiple access (OMA). Meanwhile, improved user fairness can be achieved by NOMA.

Nonetheless, the promised advantages of NOMA cannot be realized without proper resource allocation. The main resources in wireless communication systems include time, frequency, space, code and power. In NOMA systems, multiple users are accommodated in each time/frequency/code resource block (RB), forming a NOMA cluster. As a result, how to group the users into NOMA clusters and allocate the power is of significance. A large number of studies have been carried out for developing efficient power allocation (PA) algorithms in single-input single-output (SISO) scenarios with fixed user clustering. To fully reap the gain of NOMA, the design of joint PA and user clustering is required. Moreover, the study of PA under multiple-input multiple-output (MIMO) systems still remains at an incipient stage. In this dissertation, we develop novel algorithms to allocate resource for both SISO-NOMA and MIMO-NOMA systems.

More specifically, Chapter 2 compares the system capacity of MIMO-NOMA with MIMO-OMA. It is proved analytically that MIMO-NOMA outperforms MIMO-OMA in

terms of both sum channel capacity and ergodic sum capacity when there are multiple users in a cluster. Furthermore, it is demonstrated that the more users are admitted to a cluster, the lower is the achieved sum rate, which illustrates the tradeoff between the sum rate and maximum number of admitted users.

Chapter 3 addresses the PA problem for a general multi-cluster multi-user MIMO-NOMA system to maximize the system energy efficiency (EE). First, a closed-form solution is derived for the corresponding sum rate (SE) maximization problem. Then, the EE maximization problem is solved by applying non-convex fractional programming.

Chapter 4 investigates the energy-efficient joint user-RB association and PA problem for an uplink hybrid NOMA-OMA system. The considered problem requires to jointly optimize the user clustering, channel assignment and power allocation. To address this hard problem, a many-to-one bipartite graph is first constructed considering the users and RBs as the two sets of nodes. Based on swap matching, a joint user-RB association and power allocation scheme is proposed, which converges within a limited number of iterations. Moreover, for the power allocation under a given user-RB association, a low-complexity optimal PA algorithm is proposed.

Furthermore, Chapter 5 focuses on securing the confidential information of massive MIMO-NOMA networks by exploiting artificial noise (AN). An uplink training scheme is first proposed, and on this basis, the base station precodes the confidential information and injects the AN. Following this, the ergodic secrecy rate is derived for downlink transmission. Additionally, PA algorithms are proposed to maximize the SE and EE of the system.

Finally, conclusions are drawn and possible extensions to resource allocation in NOMA systems are discussed in Chapter 6.

To my wife, my parents and my brother ...

Acknowledgements

I would like to give my sincere thanks to my supervisor Prof. Octavia A. Dobre for the amazing opportunity she provided me with. Her valuable guidance, dedication, support and encouragement have pushed me far beyond my expectations. She is more of a family than a supervisor to me.

I would like to acknowledge the financial support provided by my supervisor, the Faculty of Engineering and Applied Science, and the School of Graduate Studies.

I would like to thank the staff of the Department of Electrical and Computer Engineering. I thank all my group members and lab colleagues, such as Xiang, Ranning, Yi, Shu, Deyuan, Ruiqin, Sunish, Animesh, Ibrahim, Yahia, Phong, Yemi, Ahmed Mohamed Ali, Quang, Al-Habob, Sylvester, Ali, Esraa etc. It was fun working with all of you.

I would like to acknowledge the most impactful persons in my life, my parents. I would have never been in this position without your love, care, and support.

Finally, I would like to acknowledge a lot of sincere and loving people I have been blessed with, who have strongly contributed to this success. Starting with my brother Li Zeng, for taking care of everything back home on my behalf, my father and mother in law, what can I say to express my gratitude to you, my lovely wife, without you nothing is complete, and my kid, you add a great value to everything.

Co-Authorship Statement

I, Ming Zeng, hold a principle author status for all the manuscript chapters (Chapter 2 - 5) in this dissertation. However, each manuscript is co-authored by my supervisor and co-researchers, whose contributions have facilitated the development of this work as described below.

- Paper 1 in Chapter 2: Ming Zeng, Animesh Yadav, Octavia A. Dobre, Georgios I. Tsiropoulos, and H. Vincent Poor, "Capacity Comparison Between MIMO-NOMA and MIMO-OMA With Multiple Users in a Cluster," IEEE Journal on Selected Areas in Communications, vol. 35, no. 10, pp. 2413-2424, October 2017.
  I was the primary author, with authors 2 - 4 contributing to the development of the idea and refinement of the presentation.

- Paper 2 in Chapter 3: Ming Zeng, Animesh Yadav, Octavia A. Dobre, and H. Vincent Poor, "Energy-Efficient Power Allocation for MIMO-NOMA With Multiple Users in a Cluster," IEEE Access, vol. 6, pp. 5170-5181, February 2018.
  I was the primary author, with authors 2 - 3 contributing to the development of the idea and refinement of the presentation.

- Paper 3 in Chapter 4: Ming Zeng, Animesh Yadav, Octavia A. Dobre, and H. Vincent Poor, "Energy-Efficient Joint User-RB Association and Power Allocation for Uplink Hybrid NOMA-OMA," IEEE Internet of Things Journal, vol. 6, no. 3,

pp. 5119-5131, June 2019.

I was the primary author, with authors 2 - 3 contributing to the development of the idea and refinement of the presentation.

- Paper 4 in Chapter 5: Ming Zeng, Nam-Phong Nguyen, Octavia A. Dobre, and H. Vincent Poor, "Securing Downlink Massive MIMO-NOMA Networks with Artificial Noise," IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 3, pp. 685-699, June 2019.

  I was the primary author, with the second author contributing equally to the formulation and development of the idea, and refinement of the presentation.

_____ _____

Ming Zeng Date

Table of Contents

List of Tables

List of Figures

List of Abbreviations

5G	The Fifth Generation
AN	Artificial Noise
AWGN	Additive White Gaussian Noise
BS	Base Station
CDMA	Code Division Multiple Access
CR	Cognitive Radio
CSI	Channel State Information
EE	Energy Efficiency
FD	Full Duplex
FDMA	Frequency Division Multiple Access
KKT	Karush-Kuhn-Tucker
MC	Multi-carrier
MIMO	Multiple-input Multiple-output
NOMA	Non-orthogonal Multiple Access

OFDMA	Orthogonal Frequency Division Multiple Access
OMA	Orthogonal Multiple Access
PA	Power Allocation
PLS	Physical Layer Security
PU	Primary User
QoS	Quality-of-service
RB	Resource Block
SA	Sub-carrier Assignment
SC	Single-carrier
SE	Spectral Efficiency
SIC	Successive Interference Cancellation
SISO	Single-input Single-output
SOP	Secrecy Outage Probability
SNR	Signal-to-noise Ratio
TDMA	Time Division Multiple Access

Chapter 1

Introduction and Overview

1.1 Background

Every generation of cellular networks comes with new standards, techniques and features, differentiating it from the previous one. In line with that, the next generation cellular systems, the fifth generation (5G) and Beyond, are expected to support various advanced services including multimedia applications, Internet-of-Things based applications, and vehicle-to-everything. These innovative use cases are leading the gigantic growth of mobile traffic, which is in turn introducing radio spectrum scarcity as one of the most critical challenges that 5G and Beyond should deal with.

Multiple access, one of the fundamental building blocks in wireless communication systems, has a significant impact on the utilization of the available spectrum, system throughput and latency [1]. In the cellular radio context, multiple access refers to a technique by which multiple users share a common radio resource to establish communication links with a base station (BS). Some of the widely used multiple access techniques in the past generations of cellular networks include time division multiple access (TDMA), frequency division multiple access (FDMA), code division multiple access (CDMA) and

orthogonal frequency division multiple access (OFDMA). These techniques are referred to as orthogonal multiple access (OMA); the access of users is orthogonal in nature and, ideally, the users do not interfere with one another while they share the communication channel. In these schemes, orthogonal radio resources in time-, frequency-, code-domain or their combinations are assigned to multiple users.

However, as the number of orthogonal resources is limited, the OMA systems cannot serve a large number of users, as imposed by 5G. In contrast to OMA, non-orthogonal multiple access (NOMA) allows inter-user interference in the resource allocation of users, and thus, multiple users are served using the same resource block [2–6]. To mitigate the effect of the interference, interference cancellation schemes such as successive interference cancellation (SIC) are applied [2–6]. NOMA is shown to have the potential of handling a massive number of connections while offering a superior sum capacity and user fairness [7–11]. NOMA-based cellular networks have been projected to offer diverse data-hungry applications. The notion of NOMA in 5G cellular context was initially put forward in [12] and its superior performance was demonstrated. The attractive advantages of NOMA then sparked off a substantial amount of research [2–10, 13].

One of the central research topics is resource allocation. The main resources in wireless communication systems include time, frequency, space, code and power [1]. In NOMA systems, multiple users are accommodated in each time/frequency/code resource block (RB), forming a NOMA cluster. As a result, how to group the users into NOMA clusters and allocate the power are of significance [2–6]. User clustering is a hard problem in general, owing to the inherent combinatorial nature [14, 15]. Because of this, often heuristic methods with low-complexity are adopted, rather than seeking the optimal solution [16, 17].

In terms of power allocation (PA), its role in NOMA is further enhanced, since users are multiplexed in the power domain [5]. Interference management, rate distribution,

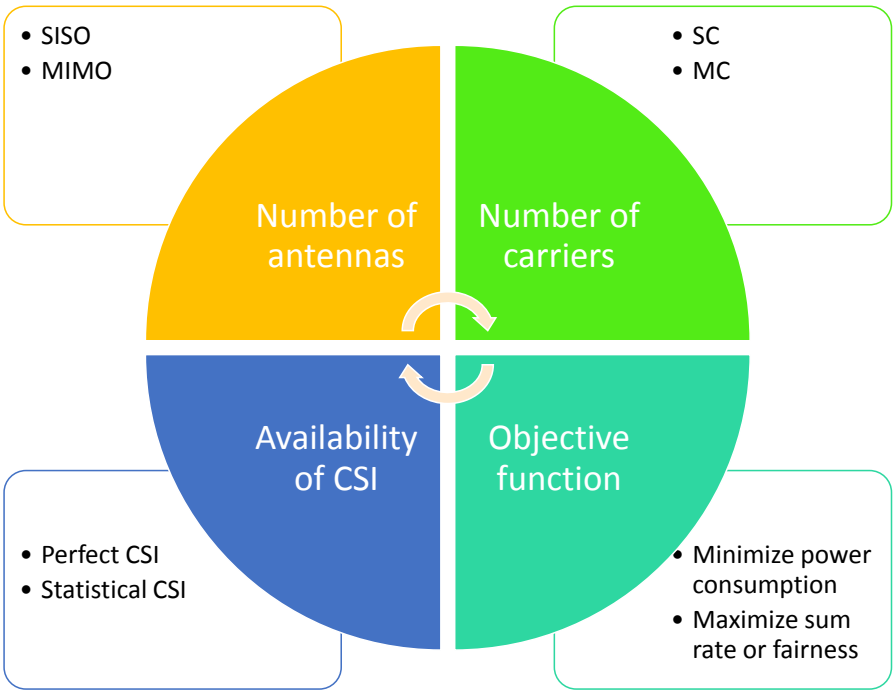


Fig. 1.1: Classification of PA strategies.

and even user admission are directly impacted by PA. Generally, PA in NOMA is determined by the users' channel conditions, availability of channel state information (CSI), quality-of-service (QoS) requirements, total power constraint and system objective. An inappropriate PA not only leads to an unfair rate distribution among users, but also causes system outage as SIC may fail. There are different PA performance metrics, e.g., the number of admitted users, sum rate, energy efficiency, user fairness, outage probability and total power consumption. Thus, PA in NOMA should aim at achieving either more admitted users, higher sum rate and energy efficiency, or a balanced fairness under minimum power consumption. A variety of PA strategies have been proposed in the literature, targeting different aspects of PA in NOMA, and a classification is provided in Fig. 1.1. We introduce PA in the following two sections: one focuses on single-input single-output (SISO) scenario, while the other deals with multiple-input multiple-output (MIMO) systems.

1.2 PA in SISO-NOMA Systems

Most of the previous work on PA has focused on SISO scenarios, including single-carrier (SC-NOMA) [11, 18–20] and multi-carrier (MC-NOMA) systems [14, 15, 21–23].

1.2.1 PA in SC-NOMA

In downlink SC-NOMA systems, the optimal PA to maximize the sum rate simply allocates all power to the user with the best channel [14, 18]. Clearly, this results in extreme unfairness among users and decreases the number of admitted users as well. To strike a balance between system throughput and user fairness, more power is allocated to the weak user in NOMA. By doing this, the strong user can remove the interference from the weak user via SIC, while its interference to the weak user remains comparatively small. Fixed PA (F-PA) is the simplest PA algorithm, and it allocates power to the users utilizing a fixed ratio based on their positions in the channel ordering [8, 13]. Since users' specific channel gains are not exploited during PA, F-PA may not meet users' various QoS requirements. To handle this, fractional transmit power control allocates power to the users inversely proportional to their channel gains powered with a decaying factor. Nonetheless, assigning the same decaying factor to all users is still suboptimal, and how to select the appropriate decaying factor to balance system throughput and user fairness is an open issue.

The availability of CSI directly impacts the PA in NOMA. Under perfect CSI, the authors in [14] show that the weighted sum rate maximization problem is convex, and obtain optimal PA via convex optimization. The max-min fairness problem is proved to be quasi-convex, and thus, optimal PA can be attained using the bisection method [11]. The energy-efficient PA problem can be formulated as a fractional problem, for which the Dinkelbach's algorithm can be applied [24, 25]. Under statistical CSI, the min-max outage

probability problem under a given SIC order is non-convex. Nonetheless, optimal PA is derived in [11], and based on this, the authors in [20] derive the optimal SIC decoding order.

The works above may not guarantee a higher throughput of NOMA over OMA for each user. To ensure this for the weak user, cognitive radio (CR)-inspired PA can be deployed, in which the weak user is viewed as a primary user in a CR network [8, 13]. However, this is achieved at the expense of the strong user as it is served only after the weak user's QoS is satisfied. To address this issue, dynamic PA is proposed in [26], which allocates power to the users such that NOMA achieves strictly higher individual user rate than OMA.

1.2.2 PA in MC-NOMA

For multi-user systems, NOMA is usually integrated with MC (MC-NOMA) to reduce the decoding complexity [27]. In MC-NOMA, a user can occupy multiple sub-carriers, and vice versa. MC-NOMA is quite suitable for 5G and Beyond since it is difficult to find continuous wide bandwidth in 5G and Beyond. When compared with OFDMA, MC-NOMA not only increases the system spectral efficiency, but also supports a larger number of users. Its performance is affected by both PA and sub-carrier assignment (SA). For downlink, the weighted sum rate maximization problem is proved to be NP-hard [14, 15]. In contrast, for uplink, the sum rate problem is shown to be convex [28], and further, an optimal and low-complexity iterative water-filling algorithm is proposed. This major difference between downlink and uplink is because the BS decodes all user signals in uplink, while each user decodes its own signal separately in downlink.

It is worth mentioning that perfect CSI is assumed in the above schemes, which might be impractical for MC-NOMA systems overloaded with exceedingly number of users. To overcome this, the resource allocation under statistical CSI should be studied. Without

perfect CSI, an explicit SIC decoding order should be derived first, as the BS cannot decide the SIC decoding order directly [22]. On this basis, PA and SA can be conducted as for the case with perfect CSI. To further increase the spectral efficiency of MC-NOMA systems, full duplex (FD) BS can be deployed, yielding a substantial throughput enhancement when compared with FD MC-OMA and half duplex MC-NOMA systems [23].

1.3 PA in MIMO-NOMA Systems

The application of MIMO to NOMA is of significance, since MIMO introduces extra spatial degrees of freedom for system performance improvement. Research on MIMO-NOMA has attracted great attention from both academia [7–10, 13] and industry [12, 29].

Compared with SISO-NOMA, the main challenge in MIMO-NOMA comes from the fact that the MIMO channel is non-degraded, i.e., users cannot be ordered based on their channel strengths in general settings. As a result, MIMO-NOMA is generally not capacity-achieving. Meanwhile, user ordering is a difficult task for MIMO-NOMA. Unlike SISO-NOMA, in which user channels are scalars and can be easily ordered, in MIMO-NOMA, user channels are in the form of vectors or matrices and cannot be ordered directly. A simple way to handle this is to order the users based only on the large-scale path loss. However, this may yield system performance degradation, since small-scale channel information is not exploited. To fully recap the spatial degrees of freedom, conceiving an appropriate beamforming/precoding design is essential for MIMO-NOMA systems. In particular, both the power domain and the angular domain should be considered in beamforming design to enhance the system spectral efficiency. There exist two popular MIMO-NOMA designs: 1) beamforming-based MIMO-NOMA design [7, 30–32], and 2) cluster-based MIMO-NOMA design [8–10, 13, 17, 33].

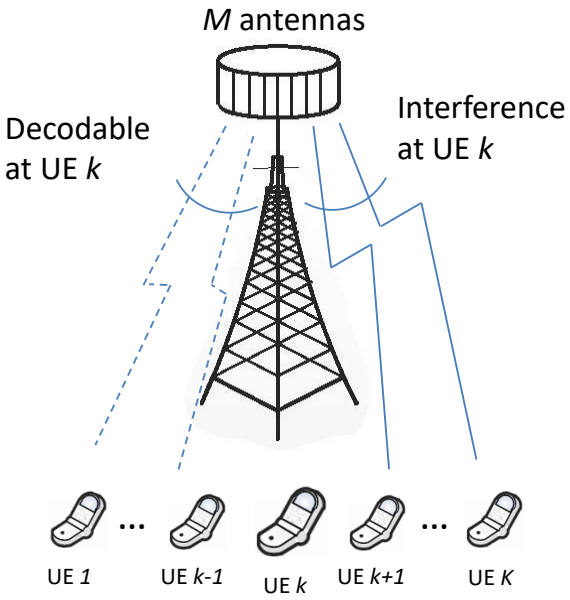


Fig. 1.2: Illustration of beamforming-based MIMO-NOMA.

1.3.1 Beamforming-based MIMO-NOMA

As shown in Fig. 1.2, in beamforming-based MIMO-NOMA, each user is assigned its own beamforming vector. However, unlike MIMO-OMA, SIC is still performed at the user side to remove the interference. To guarantee successful SIC, the following constraint should be explicitly given, i.e., the achievable rate at each user cannot exceed the minimum rate among all users which need to decode it. In [7], the ergodic sum rate maximization problem is studied for a Rayleigh fading based MIMO-NOMA systems with statistical CSI at the transmitter. Both optimal and low-complexity suboptimal PA schemes are proposed under total transmit power constraint and minimum rate constraint of the weak user. Numerical results show that the proposed NOMA schemes significantly outperform traditional OMA. In [30], a layered transmission scheme is proposed based on QR factorization. Under instantaneous CSI, an approach to maximize the sum rate of MIMO-NOMA with layered transmissions is proposed after showing that the sum rate is concave in allocated powers to multiple layers of users. Under statistical CSI, a closed-form expression for

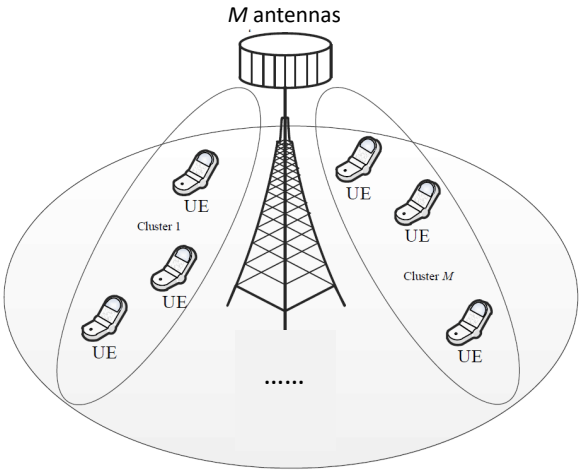


Fig. 1.3: Illustration of cluster-based MIMO-NOMA.

the average sum rate is derived, and on this basis, power is allocated using alternating optimization. In [31], optimal precoding is studied for a QoS constrained optimization problem in a MISO-NOMA network. It is shown that if the broadcast channels are quasi-degraded, the proposed optimization algorithm in combination with superposition coding and SIC can achieve system capacity. Note that [7,30,31] apply to only two users. In [32], the general scenario with multiple users is considered. The sum rate maximization is handled with the help of the minorization-maximization algorithm.

1.3.2 Cluster-based MIMO-NOMA

According to Fig. 1.3, in cluster-based MIMO-NOMA, users are first grouped into clusters based on their channels; then, the users allocated to the same cluster share a common beamforming vector [8,13]. Note that deriving the optimal user clustering often requires to exhaustively enumerate all possible user clustering combinations and is computationally prohibited. A good heuristic for this is to put users with high channel correlation and gain difference into the same cluster, since high channel correlation facilitates the beamforming design while high gain difference favors the implementation of SIC [17]. Once users are

clustered, the main task lies in how to design the appropriate beamforming vectors for each cluster. The simplest beamforming design is to use random beamforming, which is also effective in reducing the CSI feedback [13,33]. In [33], the authors propose to combine random beamforming with intra-beam SIC for downlink NOMA transmission. However, due to the existence of inter-cluster interference, different clusters are still coupled. To address this issue, the authors in [13] propose to use random beamforming at the BS, and zero-forcing detection at the user side. Then, inter-cluster interference can be removed at each user, and the MIMO-NOMA system is decomposed into a set of independent SISO-NOMA arrangements. On this basis, the impact of different power allocation strategies, namely F-PA and CR-inspired PA, on the performance of MIMO-NOMA is investigated. In addition, in [8], the authors propose to apply a signal alignment technique to the two users in the same cluster, such that they can be treated as a single user. Then, a conventional zero-forcing precoder can be used as in MIMO-OMA. The impact of F-PA and CR-inspired PA on the system performance is also investigated, as in [13].

1.4 Physical Layer Security (PLS) in NOMA Systems

Traditionally, the security issues have been handled at the higher layers using encryption approaches. However, the development of computing technologies and the tremendous growth in the number of wireless devices have surfaced the vulnerability of the conventional encryption methods [34]. As a result, PLS has been introduced as an additional protecting layer to the conventional encryption methods for securing confidential information [35]. The principle of PLS is to take advantage of the randomness of the wireless channels to restrain the illegitimate side from overhearing the legitimate users [36]. The community has shown a great interest in applying PLS to NOMA networks. In [37], the

authors investigated the secrecy outage probability (SOP) of NOMA relay networks with two types of relay, i.e., amplify-and-forward and decode-and-forward. It was found that in the high signal-to-noise ratio regime, the SOP of the considered NOMA relay network converges to a constant value. In [38], the secrecy performance of a stochastic NOMA network was considered, by modelling its users' locations using stochastic geometry. The results showed that the secrecy diversity order of the considered system is determined by that of the user pair with a poorer channel. In [39], the authors derived a closed-form solution for maximizing the secrecy sum rate of the NOMA while taking the users' quality of service requirements into consideration. In [40], the authors investigated a NOMA system in the presence of an external eavesdropper. The SOP of the considered system was derived and used to optimize the decoding order, transmission rates, and allocated power. These studies have laid the initial foundation for exploiting PLS in NOMA networks.

1.5 Motivation and Outline

Motivation:

Based on the aforementioned discussion, the following observations can be made:

- Whether MIMO-NOMA outperforms MIMO-OMA in terms of system's spectral and energy efficiencies for the general multi-cluster multi-user scenario is unclear, and requires further investigation.

- Most works on energy efficiency are focused on downlink NOMA systems. The study of energy efficiency on uplink NOMA system is of interest, since user terminals are power-constrained.

- Most works on PLS are focused on SISO-NOMA systems. The study of PLS on MIMO-NOMA systems is of significance. In MIMO-based systems, artificial noise

(AN) is often added to secure the legitimate side from malicious attempts. The role of AN in MIMO-NOMA systems is worth of investigation.

Motivated by the aforementioned observations, in this thesis we have addressed the following research problems:

Research problems:

- **P1**- Proving that MIMO-NOMA outperforms MIMO-OMA in terms of system's spectral efficiency for the general multi-cluster multi-user scenario.
- **P2**- Developing a low-complexity algorithm, which can maximize the system's energy efficiency for the general multi-cluster multi-user downlink MIMO-NOMA system.
- **P3**- Developing a joint user clustering and PA allocation algorithm to maximize the energy efficiency for an uplink hybrid NOMA-OMA system.
- **P4**- Developing a framework to secure the confidential information of MIMO-NOMA networks by exploiting AN.

Thesis structure:

Chapter 2 addresses **P1**, i.e., it proves analytically that MIMO-NOMA outperforms MIMO-OMA in terms of both sum channel capacity and ergodic sum capacity when there are multiple users in a cluster. Furthermore, it demonstrates that the more users are admitted to a cluster, the lower is the achieved sum rate, which illustrates the tradeoff between the sum rate and maximum number of admitted users.

Chapter 3 addresses **P2**, by considering the energy-efficient PA problem for a general multi-cluster multi-user MIMO-NOMA system. A closed-form solution is first derived for the corresponding sum rate maximization problem. Then, the energy efficiency maximization problem is solved by applying non-convex fractional programming.

Chapter 4 addresses **P3**, by investigating the energy-efficient joint user-RB association and PA problem for an uplink hybrid NOMA-OMA system. The considered problem requires to jointly optimize the user clustering, channel assignment and power allocation. To address this hard problem, a many-to-one bipartite graph is first constructed considering the users and RBs as the two sets of nodes. Based on swap matching, a joint user-RB association and power allocation scheme is proposed, which converges within a limited number of iterations. Moreover, for the power allocation under a given user-RB association, a low-complexity optimal PA algorithm is proposed.

Chapter 5 addresses **P4**, by studying how to secure the confidential information of MIMO-NOMA networks by exploiting AN. An uplink training scheme is first proposed, and on this basis, the base station precodes the confidential information and injects the AN. Following this, the ergodic secrecy rate is derived for downlink transmission. Finally, PA algorithms are proposed to maximize the SE and EE of the system.

1.6 Contributions

This dissertation presents the following novel contributions to the resource allocation in NOMA:

- For the first time in the literature, it proves analytically that MIMO-NOMA outperforms MIMO-OMA in terms of both sum channel capacity and ergodic sum capacity when there are multiple users in a cluster. This lays a solid foundation for the advancement of MIMO-NOMA study.

- It shows that there exists a tradeoff between the sum rate and maximum number of admitted users.

- It develops a low-complexity optimal PA algorithm to maximize the EE for a general multi-cluster multi-user downlink MIMO-NOMA system.
- It proposes an energy-efficient joint user-RB association and PA problem for an uplink hybrid NOMA-OMA system, based on matching theory.
- It proposes a framework to secure the confidential information of MIMO-NOMA networks, involving design of uplink training and injection of AN. Moreover, it proposes PA algorithms to maximize the SE and EE of the considered MIMO-NOMA system.

References

[1] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication.* Cambridge, UK: Cambridge University Press, 2005.

[2] L. Dai et al., "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[3] S. M. R. Islam et al., "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surv. Tuts.*, vol. 19, no. 2, pp. 721–742, Second quarter 2017.

[4] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[5] S. M. R. Islam, M. Zeng, and O. A. Dobre, "NOMA in 5G systems: Exciting possibilities for enhancing spectral efficiency," *IEEE 5G Tech. Focus,* vol. 1, no. 2, May 2017. [Online]. Available: 307 http://5g.ieee.org/tech-focus.

[6] S. M. R. Islam, M. Zeng, O. A. Dobre, and K. Kwak, "Resource allocation for downlink noma systems: Key techniques and open issues," *IEEE Wireless Commun. Mag.*, vol. 25, no. 2, pp. 40–47, April 2018.

[7] Q. Sun, S. Han, I. Chin-Lin, and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 4, pp. 405–408, Dec. 2015.

[8] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, Jun. 2016.

[9] 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, Aug. 2017.

[10] ——, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Select. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.

[11] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[12] Y. Saito et al., "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE Veh. Technol. Conf.*, Dresden, Germany, Jun. 2013, pp. 1–5.

[13] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[14] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Power and channel allocation for non-orthogonal multiple access in 5G systems: Tractability and computation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8580–8594, Dec. 2016.

[15] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686–7698, Nov. 2016.

[16] 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, Aug. 2016.

[17] B. Kimy et al., "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *Proc. IEEE Mil. Commun. Conf*, Nov. 2013, pp. 1278–1283.

[18] M. Zeng, G. I. Tsiropoulos, O. A. Dobre, and M. H. Ahmed, "Power allocation for cognitive radio networks employing non-orthogonal multiple access," in *Proc. IEEE Globecom*, Washington DC, USA, Dec. 2016.

[19] M. Zeng, G. I. Tsiropoulos, A. Yadav, O. A. Dobre, and M. H. Ahmed, "A two-phase power allocation scheme for CRNs employing NOMA," in *Proc. IEEE Globecom*, Singapore, Singapore, Dec. 2017, pp. 1–6.

[20] S. Shi, L. Yang, and H. Zhu, "Outage balancing in downlink nonorthogonal multiple access with statistical channel state information," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4718–4731, Jul. 2016.

[21] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for MC-NOMA systems," in *Proc. IEEE Globecom*, Washington, DC, USA, Dec 2016, pp. 1–6.

[22] Z. Wei, D. W. K. Ng, and J. Yuan, "Power-efficient resource allocation for MC-NOMA with statistical channel state information," in *Proc. IEEE Globecom*, Washington, DC, USA, Dec. 2016, pp. 1–7.

[23] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.

[24] Y. Zhang, H. M. Wang, T. X. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2852–2857, Mar. 2017.

[25] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation for uplink NOMA," in *Proc. IEEE Globecom*, Abu Dhabi, UAE, Dec. 2018, pp. 1–6.

[26] ——, "A fair individual rate comparison between MIMO-NOMA and MIMO-OMA," in *Proc. IEEE Globecom Wkshps*, Singapore, Dec 2017, pp. 1–5.

[27] ——, "Energy-efficient power allocation for hybrid multiple access systems," in *Proc. IEEE ICC Wkshps*, Kansas City, MO, USA, May 2018, pp. 1–5.

[28] M. Zeng, N. P. Nguyen, O. A. Dobre, Z. Ding, and H. V. Poor, "Spectral and energy efficient resource allocation for multi-carrier uplink NOMA systems," *IEEE Trans. Veh. Technol.*, submitted.

[29] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Trans. Commun.*, vol. 98, no. 3, pp. 403–414, Mar. 2015.

[30] J. Choi, "On the power allocation for mimo-noma systems with layered transmissions," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3226–3237, May 2016.

[31] Z. Chen, Z. Ding, P. Xu, and X. Dai, "Optimal precoding for a QoS optimization problem in two-user MISO-NOMA downlink," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1263–1266, Jun. 2016.

[32] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, "A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems," *IEEE Trans. Signal Processing*, vol. 64, no. 1, pp. 76–88, Jan. 2016.

[33] K. Higuchi and Y. Kishiyama, "Non-orthogonal access with random beamforming and intra-beam SIC for cellular MIMO downlink," in *Proc. IEEE VTC Fall*, Sep. 2013, pp. 1–5.

[34] N. Yang, L. Wang, G. Geraci, M. Elkashlan, J. Yuan, and M. D. Renzo, "Safeguarding 5G wireless communication networks using physical layer security," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 20–27, Apr. 2015.

[35] A. D. Wyner, "The wire-tap channel," *Bell Syst. Tech. J.*, vol. 54, no. 8, pp. 1355–1387, Jan. 1975.

[36] Y.-S. Shiu, S. Chang, H.-C. Wu, S. Huang, and H.-H. Chen, "Physical layer security in wireless networks: A tutorial," *IEEE Wireless Commun.*, vol. 18, no. 2, pp. 66–74, Apr. 2011.

[37] J. Chen, L. Yang, and M.-S. Alouini, "Physical layer security for cooperative NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4645–4649, May 2018.

[38] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and L. Hanzo, "Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1656 – 1672, Mar. 2017.

[39] Y. Zhang, H.-M. Wang, Q. Yang, and Z. Ding, "Secrecy sum rate maximization in non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 930–933, May 2016.

[40] B. He, A. Liu, N. Yang, and V. K. N. Lau, "On the design of secure non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2196–2206, Oct. 2017.

Chapter 2

Capacity Comparison Between MIMO-NOMA and MIMO-OMA With Multiple Users in a Cluster

2.1 Abstract

In this chapter, the performance of MIMO-NOMA is investigated when multiple users are grouped into a cluster. The superiority of MIMO-NOMA over MIMO-OMA in terms of both sum channel capacity and ergodic sum capacity is proved analytically. Furthermore, it is demonstrated that the more users are admitted to a cluster, the lower is the achieved sum rate, which illustrates the tradeoff between the sum rate and maximum number of admitted users. On this basis, a user admission scheme is proposed, which is optimal in terms of both sum rate and number of admitted users when the signal-to-interference-plus-noise ratio thresholds of the users are equal. When these thresholds are different, the proposed scheme still achieves good performance in balancing both criteria. Moreover, under certain conditions, it maximizes the number of admitted users. In addition, the

complexity of the proposed scheme is linear to the number of users per cluster. Simulation results verify the superiority of MIMO-NOMA over MIMO-OMA in terms of both sum rate and user fairness, as well as the effectiveness of the proposed user admission scheme.

2.2 Introduction

NOMA has attracted considerable attention recently due to its superior spectral efficiency [1–7]. Specifically, NOMA adopts superposition coding (SC) at the transmitter and successive interference cancellation (SIC) at the receiver. Moreover, the transmitted power allocated to the users is inversely proportional to their channel gains. This way, the user with better channel gain can handle the interference from its counterpart, while its interference to the counterpart remains comparatively small. Thus, NOMA achieves a better balance between sum rate and fairness when compared with conventional OMA scheme, in which more power is assigned to the users with better channel conditions to increase the sum rate [8].

It is of great interest to conduct comparisons between NOMA and OMA. Early works mainly focus on SISO systems. For instance, simulation results in [1] show that a larger sum rate is achieved by NOMA, whereas in [9], it is proved that NOMA strictly dominates OMA via the achievable rate region. However, no analytical proof is provided in [1] and [9]. In [10], the performance of NOMA is investigated in a cellular downlink scenario with randomly deployed users, and the developed analytical results show that NOMA can achieve superior performance in terms of ergodic sum rate. In [8], the problem of maximizing the fairness among users of a NOMA downlink system is studied in terms of data rate under full CSI and outage probability under average CSI. Simulation results verify the efficiency of NOMA, which also achieves improved fairness when compared to

time division multiple access.

Emerging research activities in future mobile wireless networks study the performance of NOMA under MIMO channels. In [11], the authors explore the two user power allocation problem of a NOMA scheme by maximizing the ergodic sum capacity of MIMO channel under the total transmit power, minimum rate requirement and partial CSI availability constraints. Optimal and lower complexity power allocation schemes are proposed, and numerical results show that MIMO-NOMA obtains a larger ergodic sum capacity when compared to MIMO-OMA. In [12, 13], Ding et al. investigate the performance of MIMO-NOMA when there are multiple clusters in the system and, through simulations, validate the superiority of MIMO-NOMA over MIMO-OMA. Specifically, [12] studies the downlink (DL) with limited feedback at the base station (BS), while [13] considers both DL and uplink with full CSI at the user side and BS. Additionally, for each cluster, multiple users can be admitted into [12], whereas [13] can only support two users performing signal alignment. However, neither [12] nor [13] provides an analytical comparison between MIMO-NOMA and MIMO-OMA in terms of sum rate. Based on the system model proposed in [12], [14] conducts the sum rate comparison between them when there are only two users in each cluster. It is shown analytically that for any rate pair achieved by MIMO-OMA, there is a power split for MIMO-NOMA whose rate pair is larger. Despite the attractiveness of the result, its main issue is that the authors use the Jensen's inequality and concavity of $\log(\cdot)$ inappropriately to obtain the upper bound sum rate for MIMO-OMA. In [15], it is shown that for a simple scenario of two users, MIMO-NOMA dominates MIMO-OMA in terms of sum rate. Furthermore, for a more practical scenario of multiple users, with two users paired into a cluster and sharing a common transmit beamforming vector, the conclusion still holds.

Most of the existing works in MIMO-NOMA focus on the case of two users in each cluster [3,11–16], which leads to a less-studied alternative in the case of multiple users [12,

17]. In order to serve more users simultaneously, it is of great significance to investigate the performance of MIMO-NOMA with multiple users per cluster. Although [12] can support multiple users per cluster, the authors focus on user pairing and power allocation for the two user case. In [17], the proposed MIMO-NOMA scheme requires only one-bit feedback, but power allocation is not addressed, and there is no theoretical comparison of the performance of MIMO-NOMA and MIMO-OMA. In this chapter, we aim to analytically compare the performance of MIMO-NOMA with MIMO-OMA in terms of the sum channel capacity and ergodic sum capacity rather than merely providing simulation results, when there are multiple users in a cluster. Furthermore, the study of the way the sum rate varies as the number of admitted users increases in each cluster is conducted. To the best of our knowledge, this chapter is the first to address this issue under MIMO-NOMA systems. Following this, optimal user admission is investigated in terms of the number of admitted users and sum rate, when the target signal-to-interference-plus-noise ratio (SINR) of each user is given. Compared with the existing works, the main contribution of this chapter lies in:

- This chapter proves analytically that MIMO-NOMA outperforms MIMO-OMA in terms of both sum channel capacity and ergodic sum capacity when there are multiple users in a cluster. It shows that for any power split in MIMO-OMA, a larger sum rate can be achieved by MIMO-NOMA via simply assigning the same power coefficient to the latter. In addition, for the case of two users per cluster, the power split that maximizes the sum rate gap between MIMO-NOMA and MIMO-OMA is derived. Meanwhile, numerical results validate that MIMO-NOMA also achieves higher user fairness than MIMO-OMA when there are two or three users in a cluster.

- This chapter demonstrates that as more users are admitted to a cluster, the sum rate decreases. This illustrates that a tradeoff has to be considered between the

sum rate and number of admitted users. On this basis, a user admission scheme is proposed, which aims to maximize the number of admitted users under given SINR thresholds. The proposed scheme is shown to be optimal when the SINR thresholds for users in the same cluster are equal. Otherwise, it achieves a good balance between the sum rate and number of admitted users. Furthermore, under certain conditions, the proposed scheme maximizes the number of admitted users. Additionally, its complexity is linear.

The rest of the chapter is organized as follows. The system model is introduced in Section 2.3. In Section 2.4, the capacity comparison between MIMO-NOMA and MIMO-OMA is conducted. The proposed user admission scheme is introduced in Section 2.5, while simulation results are shown in Section 2.6. In Section 2.7, conclusions are drawn.

2.3 System Model

A downlink multiuser MIMO system is considered in this chapter, where the BS with M antennas transmits data to multiple receivers, each with N antennas. There are a total of ML users in the system, which are randomly grouped into M clusters with L ($L \geq 2$) users per cluster. The links between the BS and users are assumed to be quasi-static independent and identically distributed (i.i.d.) fading channels. Specifically, $\mathbf{H}_{m,l} \in \mathbb{C}^{N\times M}$ and $\mathbf{n}_{m,l} \in \mathbb{C}^{N\times 1}$ respectively represent the channel matrix and the additive white Gaussian noise vector for the lth user in the mth cluster, i.e., user (m,l) ($m \in \{1,\ldots,M\}, l \in \{1,\ldots,L\}$). Additionally, $\mathbf{P} \in \mathbb{C}^{M\times M}$ denotes the precoding matrix used by the BS, while $\mathbf{v}_{m,l} \in \mathbb{C}^{N\times 1}$ denotes the detection vector for user (m,l). The precoding matrices and detection vectors are designed as follows [12]: a) $\mathbf{P} = \mathbf{I}_M$, where $\mathbf{I}_M$ denotes the $M \times M$ identity matrix; b) $|\mathbf{v}_{m,l}|^2 = 1$ and $\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_k = 0$ for any $k \neq m$, where $\mathbf{p}_k$ is the kth column of $\mathbf{P}$. The number of antennas at the user is assumed to be equal

or larger than that at the BS to ensure the feasibility of $\mathbf{v}_{m,l}$. On this basis, for user (m,l), only a scalar value $|\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2$ needs to be fed back to the BS. Moreover, the interference from the users in all the other clusters can be removed even when there are multiple users in a cluster [12].

The performance of two multiple access schemes are compared, namely, MIMO-NOMA and MIMO-OMA.

2.3.1 MIMO-NOMA

For MIMO-NOMA scheme, SC is employed at the transmitter side, i.e., the transmitted signals share the same frequency and time resources but vary in power. Thus, the signals transmitted from the BS are given by

$$\mathbf{x} = \mathbf{P}\mathbf{s}, \tag{2.1}$$

where the information-bearing vector $\mathbf{s} \in \mathbb{C}^{M\times 1}$ can be expressed as

$$\mathbf{s} = \begin{bmatrix} \sqrt{\Omega_{1,1}} s_{1,1} + \cdots + \sqrt{\Omega_{1,L}} s_{1,L} \\ \vdots \\ \sqrt{\Omega_{M,1}} s_{M,1} + \cdots + \sqrt{\Omega_{M,L}} s_{M,L} \end{bmatrix}, \tag{2.2}$$

where $s_{m,l}$ and $\Omega_{m,l}$ are the signal and the corresponding power allocation coefficient intended for user (m,l), satisfying $\sum_{l=1}^{L} \Omega_{m,l} = 1, \forall m \in \{1,\ldots,M\}$. Without loss of generality, we set the total power to 1 for the convenience of analysis.

Further, the received signal at user (m,l) is given by

$$\mathbf{y}_{m,l} = \mathbf{H}_{m,l}\mathbf{P}\mathbf{s} + \mathbf{n}_{m,l}. \tag{2.3}$$

By applying the detection vector $\mathbf{v}_{m,l}$ on the received signal, we can easily obtain

$$\mathbf{v}_{m,l}^H \mathbf{y}_{m,l} = \mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m \sum_{l=1}^{L} \sqrt{\Omega_{m,l}} s_{m,l} + \underbrace{\sum_{k=1,k\neq m}^{M} \mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_k \mathbf{s}_k}_{\text{interference from other clusters}} + \mathbf{v}_{m,l}^H \mathbf{n}_{m,l}, \quad (2.4)$$

where $\mathbf{s}_k$ denotes the kth row of $\mathbf{s}$.

Due to the constraint[1] on the detection vector, i.e., $\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_k = 0$ for any $k \neq m$, the above equation can be simplified as

$$\mathbf{v}_{m,l}^H \mathbf{y}_{m,l} = \mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m \sum_{l=1}^{L} \sqrt{\Omega_{m,l}} s_{m,l} + \mathbf{v}_{m,l}^H \mathbf{n}_{m,l}. \quad (2.5)$$

Without loss of generality, the effective channel gains are rearranged as

$$|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1} \mathbf{p}_m|^2 \geq \cdots \geq |\mathbf{v}_{m,L}^H \mathbf{H}_{m,L} \mathbf{p}_m|^2. \quad (2.6)$$

At the receiver side, SIC will be conducted by user (m, l) to remove the interference from the users with worse channel gains, i.e., $(m, l+1), \ldots, (m, L)$. At this juncture, the following lemma is helpful to understand the efficient performance of SIC at user (m, l).

Lemma 2.1. *The interference from user* $(m, k), \forall k \in \{l+1, \ldots, L\}$ *can be removed at user* (m, l).

Proof: Refer to Appendix. ■

Remark. *Lemma 2.1 shows that under the given system model, the interference from users with worse channel conditions can be removed. Consequently, the achieved data*

[1]Owing to the specific selection of $\mathbf{P}$, this constraint is further reduced to $\mathbf{v}_{m,l}^H \tilde{\mathbf{H}}_{m,l} = 0$, where $\tilde{\mathbf{H}}_{m,l} = [\mathbf{h}_{1,ml} \cdots \mathbf{h}_{m-1,ml}\, \mathbf{h}_{m+1,ml} \cdots \mathbf{h}_{M,ml}]$ and $\mathbf{h}_{i,ml}$ is the ith column of $\mathbf{H}_{m,l}$ [12]. Hence, $\mathbf{v}_{m,l}$ can be expressed as $\mathbf{U}_{m,l}\mathbf{w}_{m,l}$, where $\mathbf{U}_{m,l}$ is the matrix consisting of the left singular vectors of $\tilde{\mathbf{H}}_{m,l}$ corresponding to the non-zero singular values, and $\mathbf{w}_{m,l}$ is the maximum ratio combining vector expressed as $\mathbf{U}_{m,l}^H \mathbf{h}_{m,ml} / |\mathbf{U}_{m,l}^H \mathbf{h}_{m,ml}|$.

rate at user (m,l) *is given by*

$$R_{m,l}^{NOMA} = \log_2\left(1 + \frac{\rho\Omega_{m,l}|\mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_m|^2}{1+\rho\sum_{k=1}^{l-1}\Omega_{m,k}|\mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_m|^2}\right), \tag{2.7}$$

where $\rho = 1/\sigma_n^2$*, with* σ_n^2 *as the noise variance. We assume that the noise variance is the same for all users.*

2.3.2 MIMO-OMA

For the OMA scheme, the same power coefficients are allocated to the L users per cluster as for the case of MIMO-NOMA for the sake of comparison, i.e., $\Omega_{m,1},\ldots,\Omega_{m,L}$. In addition, the degrees of freedom (time or frequency) are split amongst the L users per cluster, i.e., user (m,l) is assigned a fraction of the degrees of freedom, denoted by $\lambda_{m,l}$, satisfying $\sum_{l=1}^{L}\lambda_{m,l} = 1$. Accordingly, the achieved data rate at user (m,l) is given by [9]

$$R_{m,l}^{\mathrm{OMA}} = \lambda_{m,l}\log_2\left(1 + \frac{\rho\Omega_{m,l}|\mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_m|^2}{\lambda_{m,l}}\right). \tag{2.8}$$

The following lemma gives the sum rate upper bound when two users are paired in a cluster.

Lemma 2.2. *The sum rate for two users* $S_{m,2}^{OMA}$ *is bounded by [15]*

$$S_{m,2}^{OMA} \leq \log_2(1 + \sum_{l=1}^{2}\rho\Omega_{m,l}|\mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_m|^2), \tag{2.9}$$

where the equality holds when

$$\lambda_{m,l} = \frac{\Omega_{m,l}|\mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_m|^2}{\sum_{k=1}^{2}\Omega_{m,k}|\mathbf{v}_{m,k}^H\mathbf{H}_{m,k}\mathbf{p}_m|^2}, l \in \{1,2\}. \tag{2.10}$$

Remark. *Lemma 2.2 gives the maximum sum rate of two users for MIMO-OMA. On this basis, the bound of the sum rate for the mth cluster can be derived, when there are L users.*

Theorem 2.1. *The sum rate in the mth cluster is upper bounded by*

$$S_{m,L}^{OMA} \leq \log_2(1 + \sum_{l=1}^{L} \rho \Omega_{m,l} |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2), \tag{2.11}$$

where the equality holds when

$$\lambda_{m,l} = \frac{\Omega_{m,l} |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2}{\sum_{k=1}^{L} \Omega_{m,k} |\mathbf{v}_{m,k}^H \mathbf{H}_{m,k} \mathbf{p}_m|^2}, l \in \{1, \ldots, L\}. \tag{2.12}$$

Proof: Refer to Appendix. ■

Remark. *Theorem 2.1 shows that once the power allocation coefficients are ascertained, the optimal allocation of degrees of freedom can be obtained accordingly to ensure that the maximum sum rate for the mth cluster $S_{m,L}^{OMA}$ is achieved.*

2.4 Capacity Comparison between MIMO-NOMA and MIMO-OMA

In this section, both sum channel capacity and ergodic sum capacity for the mth cluster achieved by MIMO-NOMA are compared to that achieved by MIMO-OMA.

2.4.1 Sum Channel Capacity

The sum rate for MIMO-OMA has already been obtained, i.e., (2.11) and (2.12). Now, the sum rate for the mth cluster in MIMO-NOMA is considered, which is $S_{m,L}^{\text{NOMA}} = \sum_{l=1}^{L} R_{m,l}^{\text{NOMA}}$, and can be easily expressed as

$$S_{m,L}^{\mathrm{NOMA}} = \sum_{l=1}^{L} \log_2 \left(1 + \frac{\rho \Omega_{m,l} |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2}{1+\rho \sum_{k=1}^{l-l} \Omega_{m,k} |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2} \right). \tag{2.13}$$

Lemma 2.3. *The lower bound of the sum rate for MIMO-NOMA is given by*

$$S_{m,L}^{NOMA} \geq \log_2(1 + \rho \sum_{l=1}^{L} \Omega_{m,l} |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2). \tag{2.14}$$

Proof: Refer to Appendix. ■

Theorem 2.2. *For any power split in MIMO-OMA, a larger sum rate can be achieved by MIMO-NOMA via assigning the same power split to the latter. In particular, when the power split is optimal for MIMO-OMA, a larger sum channel capacity can be achieved by MIMO-NOMA.*

Proof: Combining Theorem 2.1 and Lemma 2.3, i.e., (2.11) and (2.14), we obtain

$$S_{m,L}^{\mathrm{NOMA}} \geq S_{m,L}^{\mathrm{OMA}}, \tag{2.15}$$

which proves the superiority of MIMO-NOMA over MIMO-OMA in terms of sum rate for any power split.

When the power split is optimal for MIMO-OMA, the sum channel capacity, denoted as $C_{m,L}^{\mathrm{OMA}}$, is achieved if (2.12) is met. Let us assign the same power split to MIMO-NOMA and denote its sum rate as $S_{m,L}^{\prime\mathrm{NOMA}}$. We also denote the sum channel capacity for MIMO-NOMA as $C_{m,L}^{\mathrm{NOMA}}$, which satisfies $C_{m,L}^{\mathrm{NOMA}} \geq S_{m,L}^{\prime\mathrm{NOMA}}$. Thus, we have

$$C_{m,L}^{\mathrm{NOMA}} \geq S_{m,L}^{\prime\mathrm{NOMA}} \geq C_{m,L}^{\mathrm{OMA}}, \tag{2.16}$$

where the second inequality comes from (2.15). Therefore, MIMO-NOMA achieves a larger sum channel capacity than MIMO-OMA. ■

In summary, it is proved analytically that for any instantaneous channel gain $\mathbf{H}_{m,l}$ ($m \in \{1,\ldots,M\}, l \in \{1,\ldots,L\}$), given the power split in MIMO-OMA, a larger sum rate can be achieved by MIMO-NOMA via simply allocating the same power split to the latter. Note that there is no constraint on the value of the power split, which means that the conclusion is true for any power split. Therefore, we can conclude that even when there are multiple users per cluster, MIMO-NOMA strictly outperforms MIMO-OMA in terms of the sum rate under any instantaneous channel gain $\mathbf{H}_{m,l}$ and any power split. On this basis, it is shown that MIMO-NOMA also achieves a larger sum channel capacity than MIMO-OMA.

Furthermore, when there are only two users per cluster, the following lemma provides the power allocation coefficient such that the gap between the sum rate of MIMO-NOMA and MIMO-OMA is maximized.

Lemma 2.4. *The sum rate gap for two users between MIMO-NOMA and MIMO-OMA is maximized, when the following equation is satisfied*

$$\Omega_{m,1} = \frac{\sqrt{\rho|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\mathbf{p}_m|^2 + 1} - 1}{\rho|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\mathbf{p}_m|^2}. \tag{2.17}$$

Proof: According to (2.9) and (2.13), the sum rate gap between MIMO-NOMA and MIMO-OMA is given by

$$\begin{aligned} \triangle S_{m,2} =& \log_2\{1 + \rho\Omega_{m,1}|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\mathbf{p}_m|^2\} \\ &+ \log_2\left\{1 + \frac{\rho\Omega_{m,2}|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\mathbf{p}_m|^2}{1 + \rho\Omega_{m,1}|\mathbf{v}_{m,2}^H \mathbf{H}_{m,2}\mathbf{p}_m|^2}\right\} \\ &- \log_2(1 + \sum_{l=1}^{2} \rho\Omega_{m,l}|\mathbf{v}_{m,l}^H \mathbf{H}_{m,l}\mathbf{p}_m|^2). \end{aligned} \tag{2.18}$$

After replacing $\Omega_{m,2}$ with $1 - \Omega_{m,1}$, the only variable is $\Omega_{m,1}$. It can be easily proved that when (2.17) is satisfied, $\frac{\partial \triangle S_{m,2}}{\partial \Omega_{m,1}} = 0$. Moreover, $\frac{\partial \triangle S_{m,2}}{\partial \Omega_{m,1}} > 0$ when $\Omega_{m,1} <$

$\frac{\sqrt{\rho|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2+1}-1}{\rho|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2}$, and $\frac{\partial\triangle S_{m,2}}{\partial\Omega_{m,1}} < 0$, otherwise. Therefore, the sum rate gap is maximized when (2.17) holds. In addition, since $\rho|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2 > 0$, it can be easily proven that $0 < \frac{\sqrt{\rho|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2+1}-1}{\rho|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2} < 1$, which fits the range of $\Omega_{m,1}$. ■

Accordingly, for the two user case, we can calculate the maximum sum rate gap between MIMO-NOMA and MIMO-OMA by substituting the value of $\Omega_{m,1}$ from (2.17) into (2.18).

Remark. *It is somewhat surprising that the power coefficient maximizing the sum rate gap is only determined by the channel of the first user. Moreover, according to (2.17), it can be easily verified that* $\Omega_{m,1}$ *declines with* $\rho|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2$. *Specifically, when* $\rho|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2 \to 0$, $\Omega_{m,1} \to 0.5$, *and* $\rho|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2 \to \infty$, $\Omega_{m,1} \to 0$. *Thus, it can be further concluded that* $\Omega_{m,1} < 0.5$ *for any value of* $\rho|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2$. *This is consistent with the concept of NOMA, in which a larger proportion of power should be allocated to the user with worse channel condition.*

2.4.2 Ergodic Sum Capacity

Corollary 1. *For any power split in MIMO-OMA, a larger ergodic sum rate can be achieved by MIMO-NOMA via assigning the same power split to the latter. In particular, when the power split is optimal for MIMO-OMA, a larger ergodic sum capacity can be achieved by MIMO-NOMA.*

Proof: As shown in the previous section, MIMO-NOMA strictly outperforms MIMO-OMA in terms of sum rate under any instantaneous channel gains of $\mathbf{H}_{m,l}$. By applying the expectation operator, it is straightforward to claim that the ergodic sum rate of MIMO-NOMA is always larger than that of MIMO-OMA. Likewise, it is easy to verify that the ergodic sum capacity of MIMO-NOMA is always larger than that of MIMO-OMA. Additionally, it is worth noticing that the conclusions hold regardless of

the distribution of $\mathbf{H}_{m,l}$. ■

To summarize, the same conclusion as for the sum channel capacity holds true for the ergodic sum capacity. Thus, even for the case of multiple users per cluster, MIMO-NOMA strictly outperforms MIMO-OMA in terms of both sum channel capacity and ergodic sum capacity.

2.5 User Admission

Analytical results obtained in the previous section validate that MIMO-NOMA strictly outperforms MIMO-OMA in terms of both sum rate and ergodic sum rate, even when there are multiple users in a cluster. Does this mean we should group a large number of users in a cluster to increase the system capacity in terms of the number of users? Clearly, SIC at the receiver becomes increasingly complicated when more users are included in a cluster, which limits the practical number of users per cluster. Furthermore, the study of how the sum rate varies with the number of admitted users is of interest, which we explore in the following section.

2.5.1 Sum Rate versus Number of Users

Here the MIMO-NOMA sum rate between the case of l and $l+1$ users in the mth cluster is compared. For notational simplicity, the index of the cluster, m, and the NOMA superscript are omitted. The power allocation coefficients for 1-to-l and 1-to-$(l+1)$ users are denoted as $\Omega_1, \ldots, \Omega_l$ and $\Theta_1, \ldots, \Theta_{l+1}$ respectively, satisfying $\sum_{k=1}^{l} \Omega_k = \sum_{k=1}^{l+1} \Theta_k = 1$, and $\Omega_k \geq \Theta_k, \forall k \in \{1, \ldots, l\}$. Additionally, we set $\Xi_k = \rho|\mathbf{v}_k^H \mathbf{H}_k \mathbf{p}|^2, k \in \{1, \ldots, l+1\}$ for notational simplicity, and the effective channel of the users follow the order in (3.6), i.e., $\Xi_1 \geq \cdots \geq \Xi_{l+1}$.

According to (3.7), the sum rate up to l users can be easily re-written as

$$\begin{aligned} S^{(l)} &= \sum_{k=1}^{l} R_k^{(l)} \\ &= \log_2(1+\Omega_1\Xi_1) + \sum_{k=2}^{l} \log_2\left(\frac{1+\sum_{i=1}^{k}\Omega_i\Xi_k}{1+\sum_{i=1}^{k-1}\Omega_i\Xi_k}\right), \end{aligned} \tag{2.19}$$

where $R_k^{(l)}$ denotes the rate of the kth user for the case of l users in total.

Likewise, the sum rate for the $l+1$ users can be expressed as

$$\begin{aligned} S^{(l+1)} =& \sum_{k=1}^{l+1} R_k^{(l+1)} \\ =& \log_2(1+\Theta_1\Xi_1) + \sum_{k=2}^{l} \log_2 \frac{1+\sum_{i=1}^{k}\Theta_i\Xi_k}{1+\sum_{i=1}^{k-1}\Theta_i\Xi_k} \\ &+ \log_2 \frac{1+\Xi_{l+1}}{1+\sum_{i=1}^{l}\Theta_i\Xi_{l+1}}, \end{aligned} \tag{2.20}$$

where $R_k^{(l+1)}$ denotes the rate of the kth user for the case of $l+1$ users in total.

Combining (2.19) and (2.20), the difference between the two sum rates, denoted by $\Lambda = S^{(l+1)} - S^{(l)}$, can be expressed as

$$
\begin{aligned}
\Lambda &= \log_2 \frac{1+\Theta_1\Xi_1}{1+\Omega_1\Xi_1} + \log_2 \frac{1+\Xi_{l+1}}{1+\sum_{i=1}^{l}\Theta_i\Xi_{l+1}} \\
&\quad + \sum_{k=2}^{l} \log_2 \frac{1+\sum_{i=1}^{k}\Theta_i\Xi_k}{1+\sum_{i=1}^{k-1}\Theta_i\Xi_k} \times \frac{1+\sum_{i=1}^{k-1}\Omega_i\Xi_k}{1+\sum_{i=1}^{k}\Omega_i\Xi_k} \\
&= \log_2 \frac{1+\Theta_1\Xi_1}{1+\Omega_1\Xi_1} + \log_2 \frac{1+\Xi_{l+1}}{1+\sum_{i=1}^{l}\Theta_i\Xi_{l+1}} \\
&\quad + \sum_{k=2}^{l} \log_2 \frac{1+\sum_{i=1}^{k}\Theta_i\Xi_k}{1+\sum_{i=1}^{k}\Omega_i\Xi_k} \times \frac{1+\sum_{i=1}^{k-1}\Omega_i\Xi_k}{1+\sum_{i=1}^{k-1}\Theta_i\Xi_k} \\
&= \log_2 \left\{ \frac{1+\Theta_1\Xi_1}{1+\Omega_1\Xi_1} \times \frac{1+\Xi_{l+1}}{1+\sum_{i=1}^{l}\Theta_i\Xi_{l+1}} \right. \\
&\quad \left. \times \prod_{k=2}^{l} \frac{1+\sum_{i=1}^{k}\Theta_i\Xi_k}{1+\sum_{i=1}^{k}\Omega_i\Xi_k} \times \frac{1+\sum_{i=1}^{k-1}\Omega_i\Xi_k}{1+\sum_{i=1}^{k-1}\Theta_i\Xi_k} \right\} \\
&= \log_2 \left\{ \underbrace{\frac{1+\Theta_1\Xi_1}{1+\Omega_1\Xi_1} \times \frac{1+\Omega_1\Xi_2}{1+\Theta_1\Xi_2}}_{\Lambda_1} \right. \\
&\quad \times \underbrace{\prod_{k=2}^{l-1} \frac{1+\sum_{i=1}^{k}\Theta_i\Xi_k}{1+\sum_{i=1}^{k}\Omega_i\Xi_k} \times \frac{1+\sum_{i=1}^{k}\Omega_i\Xi_{k+1}}{1+\sum_{i=1}^{k}\Theta_i\Xi_{k+1}}}_{\Lambda_2} \\
&\quad \left. \times \underbrace{\frac{1+\sum_{i=1}^{l}\Theta_i\Xi_l}{1+\sum_{i=1}^{l}\Omega_i\Xi_l} \times \frac{1+\Xi_{l+1}}{1+\sum_{i=1}^{l}\Theta_i\Xi_{l+1}}}_{\Lambda_3} \right\}.
\end{aligned}
\tag{2.21}
$$

First, let us consider Λ_1, which is given by

$$
\Lambda_1 = \frac{1+\Theta_1\Xi_1+\Omega_1\Xi_2+\Theta_1\Xi_1\Omega_1\Xi_2}{1+\Omega_1\Xi_1+\Theta_1\Xi_2+\Omega_1\Xi_1\Theta_1\Xi_2}. \tag{2.22}
$$

Due to $(\Xi_1-\Xi_2)(\Theta_1-\Omega_1) \leq 0$, it can be easily shown that $\Lambda_1 \leq 1$.

Likewise, the same method for Λ_2 can be applied. Indeed, owing to $\sum_{i=1}^{k}(\Theta_i-\Omega_i)(\Xi_k-\Xi_{k+1}) \leq 0$, it can be easily verified that each element in Λ_2 does not exceed 1. Thus, it is obtained $\Lambda_2 \leq 1$.

As for Λ_3, by applying $\sum_{i=1}^{l} \Omega_i = 1$, we have

$$\Lambda_3 = \frac{1 + \sum_{i=1}^{l} \Theta_i \Xi_l + \Xi_{l+1} + \sum_{i=1}^{l} \Theta_i \Xi_l \Xi_{l+1}}{1 + \sum_{i=1}^{l} \Theta_i \Xi_{l+1} + \Xi_l + \sum_{i=1}^{l} \Theta_i \Xi_l \Xi_{l+1}}. \tag{2.23}$$

As $(\Xi_l - \Xi_{l+1})(\sum_{i=1}^{l} \Theta_i - 1) \leq 0$, then $\Lambda_3 \leq 1$. By combining the results for Λ_1, Λ_2 and Λ_3 in (2.21), it leads to $\Lambda \leq 0$.

To conclude, the more users are admitted, the lower the sum rate is obtained. This requires further consideration of the tradeoff between the sum rate and number of admitted users. We will thus consider the problem of maximizing the user admission when the users SINR thresholds are given.

2.5.2 Proposed User Admission Scheme

The SINR thresholds of the L users in the mth cluster are denoted as $\Gamma_1, \ldots, \Gamma_L$. In addition, the maximum number of admitted users is represented as $l, l \in \{0, 1, \ldots, L\}$. Further, the l admitted users are denoted as $a_1, a_2, \ldots, a_l$. Accordingly, the problem can be formulated as

$$\max_{\Omega} \quad l \tag{2.24a}$$

$$\text{s.t.} \quad \gamma_k \geq \Gamma_k, \qquad k \in \{a_1, a_2, \ldots, a_l\} \tag{2.24b}$$

$$\sum_{k=a_1}^{a_l} \Omega_k \leq 1, \tag{2.24c}$$

where $\mathbf{\Omega} = [\Omega_1, \ldots, \Omega_L]$ is the vector whose elements are the power allocation coefficients, and γ_k is the SINR of the kth admitted user, given by

$$\gamma_k = \frac{\rho \Omega_k |\mathbf{v}_k^H \mathbf{H}_k \mathbf{p}|^2}{1 + \rho \sum_{i=1}^{k-1} \Omega_i |\mathbf{v}_k^H \mathbf{H}_k \mathbf{p}|^2}. \tag{2.25}$$

By combining (2.24b) and (2.25), we have

$$\Omega_k \geqslant \Gamma_k \sum_{i=1}^{k-1} \Omega_i + \frac{\Gamma_k}{\rho |\mathbf{v}_k^H \mathbf{H}_k \mathbf{p}|^2}, \tag{2.26}$$

where variables are only $\sum_{i=1}^{k-1} \Omega_i$, since the other parameters, i.e., ρ, Γ_k, and $|\mathbf{v}_k^H \mathbf{H}_k \mathbf{p}|^2$, are known at the BS. Therefore, if the power coefficient among users is allocated in an ascending order, i.e., from the 1st user to the Lth user sequentially, we can obtain the power coefficient for the kth user easily, since $\sum_{i=1}^{k-1} \Omega_i$ is already known. Specifically, the power coefficient for the 1st user is calculated as

$$\Omega_1 = \frac{\Gamma_1}{\rho |\mathbf{v}_1^H \mathbf{H}_1 \mathbf{p}|^2}. \tag{2.27}$$

Sequentially and iteratively, when the power coefficient of the 1st user is known, it is employed to allocate the power coefficient to the 2nd user. According to (2.26), we have

$$\Omega_2 = \Gamma_2 \Omega_1 + \frac{\Gamma_2}{\rho |\mathbf{v}_2^H \mathbf{H}_2 \mathbf{p}|^2}. \tag{2.28}$$

Likewise, the power coefficient for the kth user can be expressed as

$$\Omega_k = \Gamma_k \sum_{i=1}^{k-1} \Omega_i + \frac{\Gamma_k}{\rho |\mathbf{v}_k^H \mathbf{H}_k \mathbf{p}|^2}. \tag{2.29}$$

Obviously, power allocation for all users can be obtained according to (2.29). However, it should be noted that the total power constraint has not been considered yet during the user admission process above. Thus, when calculating the power coefficient for the kth user, we also need to ensure that the total power assigned to users, $\sum_{i=1}^{k} \Omega_i$, does not exceed 1. This is obtained by comparing $\Gamma_k \sum_{i=1}^{k-1} \Omega_i + \frac{\Gamma_k}{\rho |\mathbf{v}_k^H \mathbf{H}_k \mathbf{p}|^2}$ with $1 - \sum_{i=1}^{k-1} \Omega_i$ during each allocation phase. Whenever $\Gamma_k \sum_{i=1}^{k-1} \Omega_k + \frac{\Gamma_k}{\rho |\mathbf{v}_k^H \mathbf{H}_k \mathbf{p}|^2} < 1 - \sum_{i=1}^{k-1} \Omega_i$, it implies that there is not enough power left to be assigned to the kth user to satisfy its SINR

requirement. Therefore, the user admission process stops and the allocated power for the kth user is zero. Evidently, the same holds for $\{k+1,\ldots,L\}$ users, i.e., $\Omega_i = 0, i \in \{k,\ldots,L\}$. The admitted users are 1st user, 2nd user, ..., $(k-1)$th user, with the allocated power coefficient given by (2.29).

As for the optimality of the proposed user admission scheme, the following theorem and corollary provide the results.

Theorem 2.3. *The proposed scheme maximizes the number of admitted users when the SINR thresholds of the users satisfy the following conditions:*

$$\frac{\Gamma_1}{|\mathbf{v}_1^H \mathbf{H}_1 \mathbf{p}|^2} \leq \cdots \leq \frac{\Gamma_l}{|\mathbf{v}_l^H \mathbf{H}_l \mathbf{p}|^2} \tag{2.30a}$$

$$\Gamma_m \leq \Gamma_n, \forall m \in \{1,\ldots,l\}, n \in \{l+1,\ldots,L\}, \tag{2.30b}$$

where l represents the total number of admitted users under the proposed scheme.

Proof: Refer to Appendix. ■

Corollary 2. *The proposed user admission scheme is optimal in terms of both sum rate and number of admitted users when the SINR thresholds of the users are equal.*

Proof: According to the channel ordering, namely 2.6, it is easy to verify that $\Gamma_k = \Gamma, k \in \{1,\ldots,L\}$ satisfies both (2.30a) and (2.30b). Thus, one can conclude that the proposed user admission scheme is optimal in terms of the number of admitted users based on Theorem 2.3. In addition, since the SINR thresholds of the users are equal, maximizing the number of admitted users also leads to the maximization of the sum rate.

■

Remark. *When the SINR thresholds of the users are different, the proposed scheme still achieves good performance in balancing the tradeoff between sum rate and number of*

admitted users. Specifically, when (3.33a) and (3.33b) are met, the proposed scheme maximizes the number of admitted users, although the sum rate may be suboptimal. On the other hand, when (3.33a) is met, but (3.33b) is violated, namely, the SINR thresholds of the admitted users are higher than that of the remaining users, the proposed scheme may be suboptimal in terms of the number of admitted users, while the sum rate is still high due to two reasons: a) the admitted users have higher SINR thresholds; b) as less users are admitted, less interference among users is introduced; therefore, an increased sum rate is obtained.

In addition, the computational complexity of the proposed user admission scheme is only linear to the number of users per cluster.

Proof: For the proposed scheme, the user admission is carried out sequentially from the 1st user to the Lth user, and for each user admission process, a constant term of operations, i.e., $O(1)$,[2] is required. In all, the computational complexity is only linear to the number of users per cluster, i.e., $O(L)$. ■

2.6 Simulation Results

In this section, simulation results are presented to verify the performance of MIMO-NOMA over MIMO-OMA, and validate the accuracy of the developed theoretical results. The parameters used in the simulations are listed in Table 2.1.

Fig. 2.1 compares the sum rate of MIMO-NOMA and MIMO-OMA in two cases: with two users and three users per cluster, respectively. The total power is set to 35 dBm in simulations, and $\Omega_{m,1}$ denotes the power coefficient for the first user. For the case of two users, the remaining power is allocated to the second user. For three users, the remaining

[2]For the kth user, the calculation of $\sum_{i=1}^{k-1}\Omega_i$ seems to require $k-1$ operations. However, if we set $S_p=\sum_{i=1}^{k-1}\Omega_i$, S_p can be updated through $S_p=S_p+\Omega_k$, and only one operation is needed. Thus, according to (2.29), only 5 operations ($2\,'+'$, $2\,'\times'$, and $1\,'/'$) are needed to obtain Ω_k.

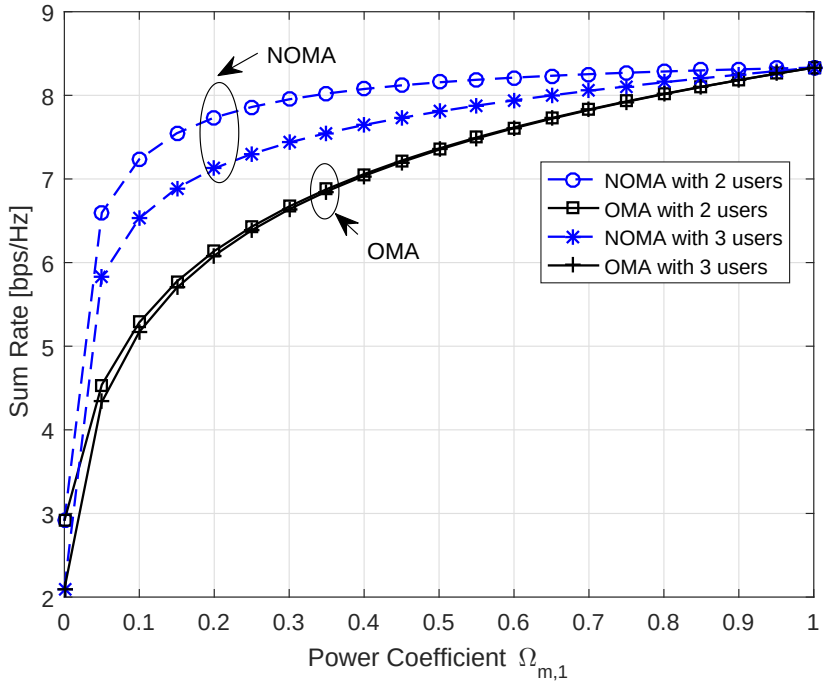


Fig. 2.1: Sum rate achieved by MIMO-NOMA and MIMO-OMA as the power coefficient varies.

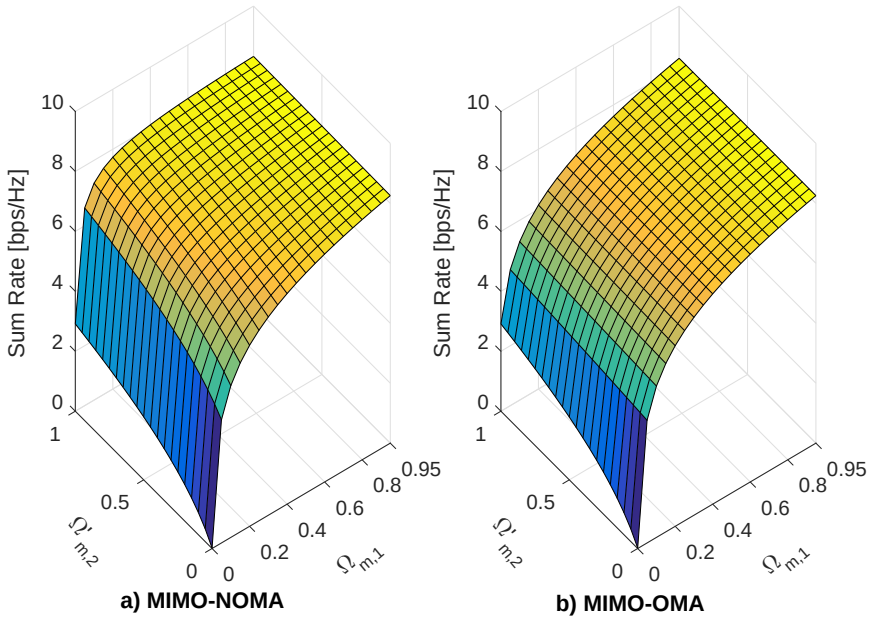


Fig. 2.2: Sum rate achieved by: a) MIMO-NOMA; b) MIMO-OMA for 3 users as the power coefficients vary.

Table 2.1: Simulation Parameters.

Parameters	**Value**
Number of antennas	$M = 3, N = 3$
Channel bandwidth	10 [MHz]
Thermal noise density	-174 [dBm]
Path-loss model	$114 + 38 \log_{10}(d)$, d in kilometer

power is equally divided between the second and third user. Note that the scenario that the remaining power is arbitrarily divided between the second and third user is shown in Fig. 2.2. Clearly, the sum rate of both MIMO-NOMA and MIMO-OMA in two cases increases with $\Omega_{m,1}$, which is due to the fact that more power is allocated to the user with better channel gain. Specifically, when $\Omega_{m,1} = 0$, for the two user case, the same sum rate is achieved for both MIMO-NOMA and MIMO-OMA, since only the second user is being served. On the other hand, for the three users case, MIMO-NOMA is slightly larger than MIMO-OMA, since two users are being served. In contrast, when $\Omega_{m,1} = 1$, the sum rate of both MIMO-NOMA and MIMO-OMA in two cases is the same since only the first user is served. In addition, for any other power split, MIMO-NOMA outperforms MIMO-OMA for both cases, which coincides with our result that MIMO-NOMA always has a larger sum rate than MIMO-OMA, even when there are multiple users in a cluster. Furthermore, for MIMO-NOMA, the two user case always has a larger sum rate when compared with the three users case, which matches the finding that when more users are admitted into a cluster, a lower sum rate is obtained.

Further, Fig. 2.2 generalizes the case for three users from Fig. 2.1, since now an arbitrary power split is provided for all three users. Thus, a three-dimensional figure is displayed, in which the y-axis scaled by $1 - \Omega_{m,1}$ represents the power coefficient of the second user, i.e., $\Omega_{m,2} = \Omega'_{m,2}(1 - \Omega_{m,1})$.[3] Additionally, the remaining power is allocated to the third user. For both MIMO-NOMA and MIMO-OMA, the sum rate increases sig-

[3]Note that in Fig. 2.2, $\Omega_{m,1}$ does not reach 1. The case of $\Omega_{m,1} = 1$ can be seen in Fig. 2.1, when the sum rates for NOMA and OMA are the same.

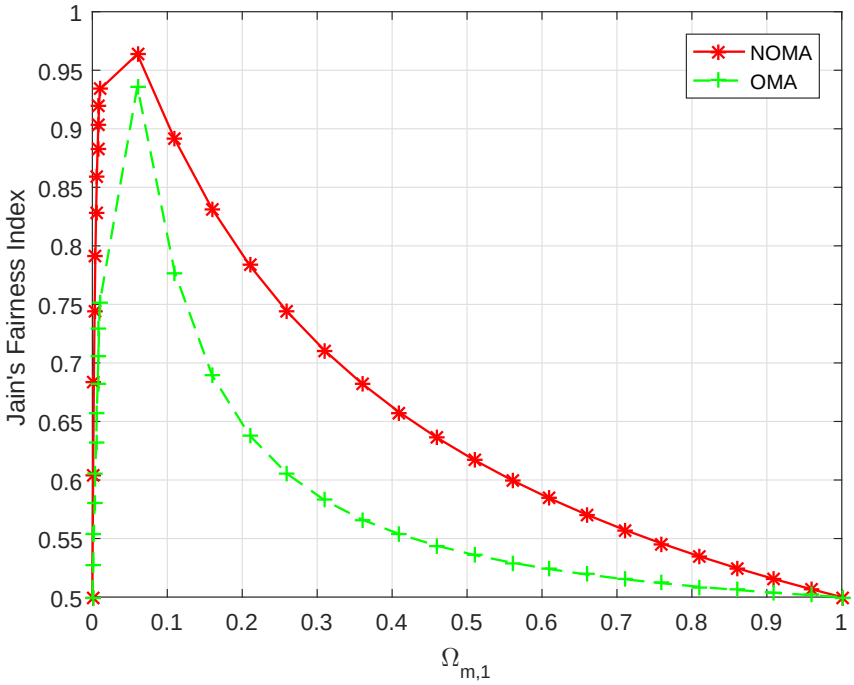


Fig. 2.3: Fairness comparison between MIMO-NOMA and MIMO-OMA for two users as the power coefficient varies.

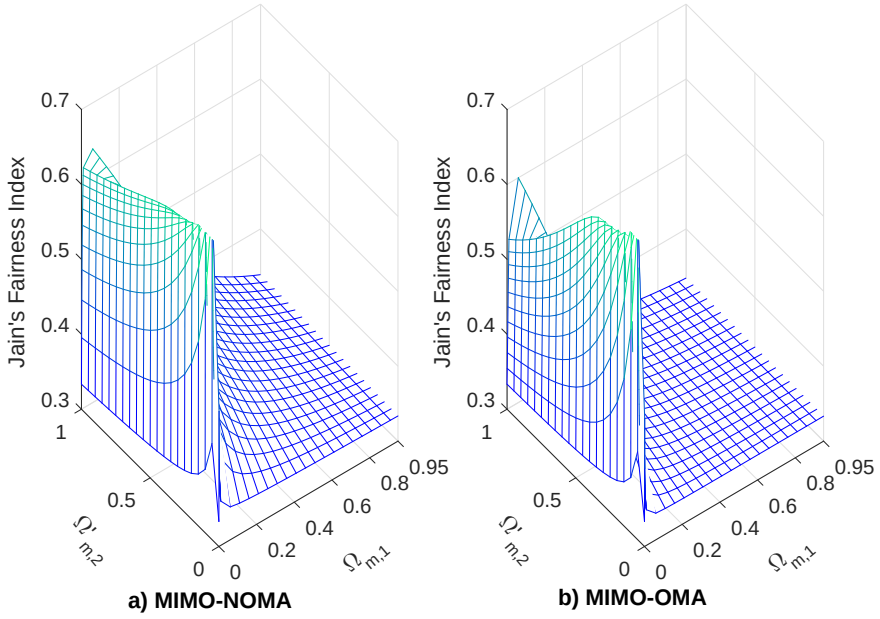


Fig. 2.4: Fairness comparison between MIMO-NOMA and MIMO-OMA for three users as the power coefficients vary.

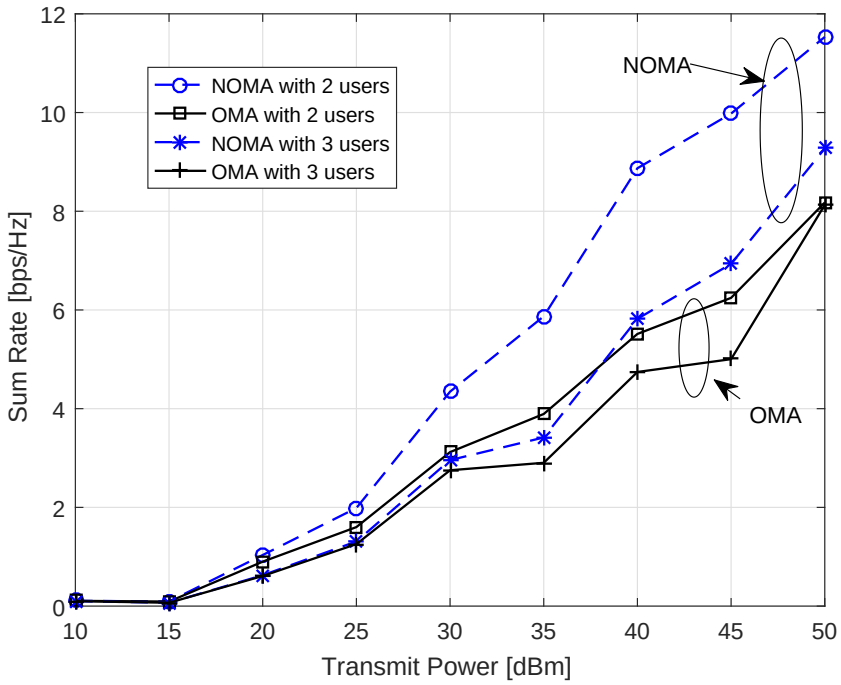


Fig. 2.5: Sum rate for MIMO-NOMA and MIMO-OMA vs. the transmit power.

nificantly with $\Omega_{m,1}$. Meanwhile, when $\Omega_{m,1}$ is fixed, both sum rates grow gradually with $\Omega_{m,2}$. These again illustrate that when more power is allocated to the user with better channel, a higher sum rate is achieved. On the other hand, when comparing Figs. 2.2a) and 2.2b), it can be seen that MIMO-NOMA always obtains a higher sum rate than MIMO-OMA for any power split among the users, which is in accordance with Theorem 2.2. Indeed, the maximum gap between MIMO-NOMA and MIMO-OMA is 2.04 bps/Hz, which is obtained at the point with $\Omega_{m,1} = 0.05, \Omega_{m,2} = 0.95$. In this case, only two users are admitted, and this can be explained by the fact that the two user case has a larger sum rate, which is likely to lead to a larger gap. For the two user case, the power allocation coefficients are consistent with the conclusion of Lemma 2.4, since during the simulation, $\rho|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\mathbf{p}_m|^2 = 321$, and thus we have $\Omega_{m,1} = 0.053$, which is close to 0.05.

Figs. 2.3 and 2.4 compare the Jain's fairness index (JFI) [16] of MIMO-NOMA and MIMO-OMA when there are two and three users in a cluster, respectively. Note that

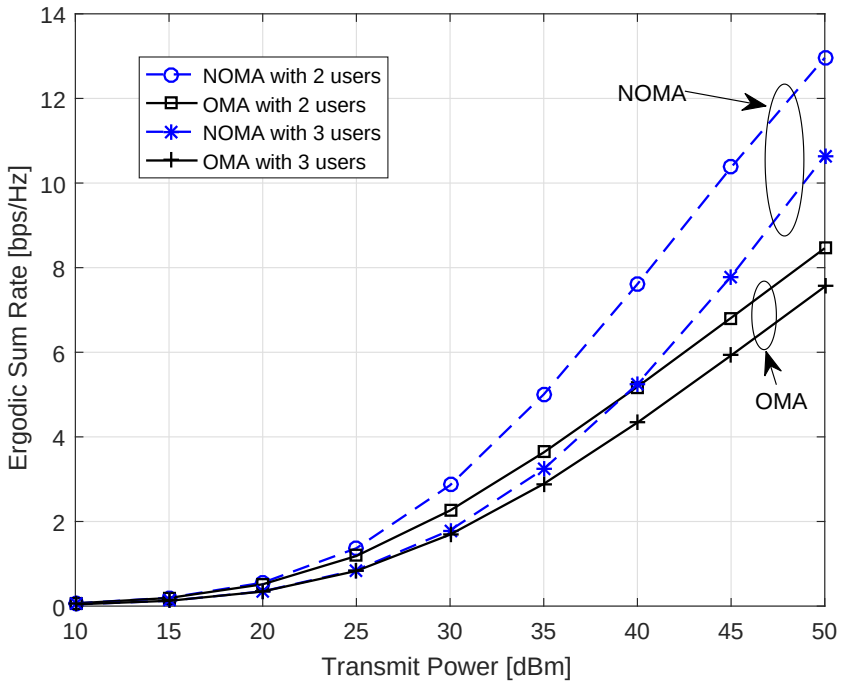


Fig. 2.6: Ergodic sum rate for MIMO-NOMA and MIMO-OMA vs. the transmit power.

$\Omega'_{m,2}$ has the same meaning as in Fig. 2.2. For both MIMO-NOMA and MIMO-OMA, for the two users case, the JFI first increases with the power coefficient to the first user ($\Omega_{m,1}$). After a certain point, i.e., around 0.1, the JFI decreases as $\Omega_{m,1}$ grows. This trend is expected, as when $\Omega_{m,1}$ is small, increasing its value leads to a more balanced rate distribution between the two users. After the point where the data rate of the first user reaches that of the second user, increasing $\Omega_{m,1}$ results in less fair rate distribution. For the three user case, as shown in Fig. 2.4, the JFI exhibits the same trend as $\Omega_{m,1}$ varies. When $\Omega_{m,1}$ is fixed, the relationship between JFI and $\Omega_{m,2}$ is more complex, and depends on the specific value of $\Omega_{m,1}$. In all, it can be seen that MIMO-NOMA dominates MIMO-OMA in both cases, which validates that MIMO-NOMA exhibits better fairness when compared with MIMO-OMA.

Figs. 2.5 and 2.6 respectively investigate the sum rate and ergodic sum rate variation with the transmit power for MIMO-NOMA and MIMO-OMA. Although there exists

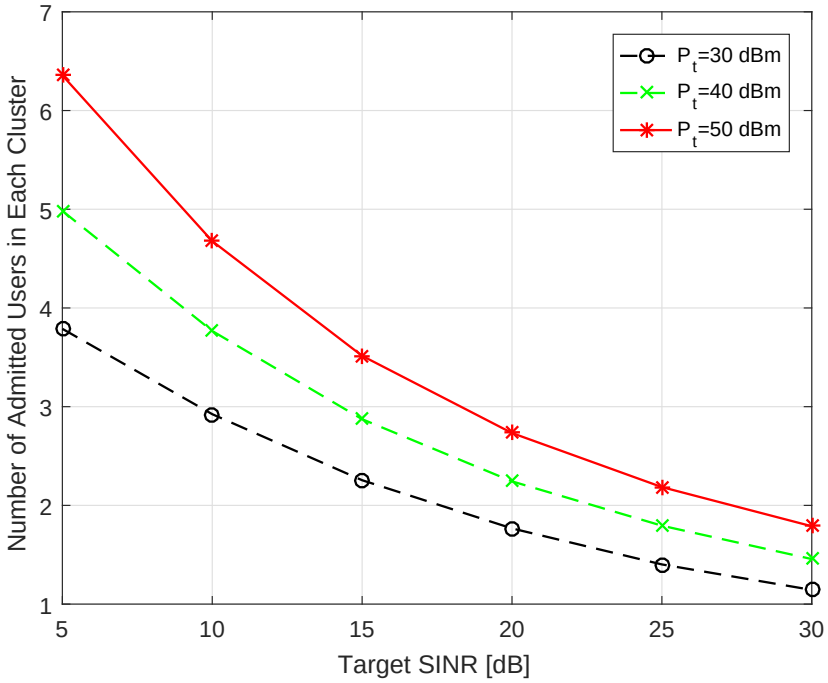


Fig. 2.7: Number of admitted users vs. target SINR.

some fluctuation in Fig. 2.5, due to the variation of the wireless channel, it is still quite clear that the sum rate of both MIMO-NOMA and MIMO-OMA grows with the transmit power. This trend becomes more obvious in Fig. 2.6, since the ergodic operation reduces the fluctuation of the channel. Moreover, in both two and three user cases, the sum rate and ergodic sum rate of MIMO-NOMA is larger than that of MIMO-OMA, which further validates our finding in Theorem 2.2. Meanwhile, as for MIMO-NOMA, the two user case always has a larger sum rate and ergodic sum rate than the three users case, which also verifies our point that as the number of admitted users increases in a cluster, the sum rate decreases.

In Figs. 2.7, 2.8 and 2.9, we focus on the performance of the proposed user admission scheme. As shown in Fig. 2.7, the number of admitted users per cluster declines with the target SINR regardless of the transmit power level. This can be easily explained by the fact that as the target SINR increases, more power is needed to satisfy each admitted

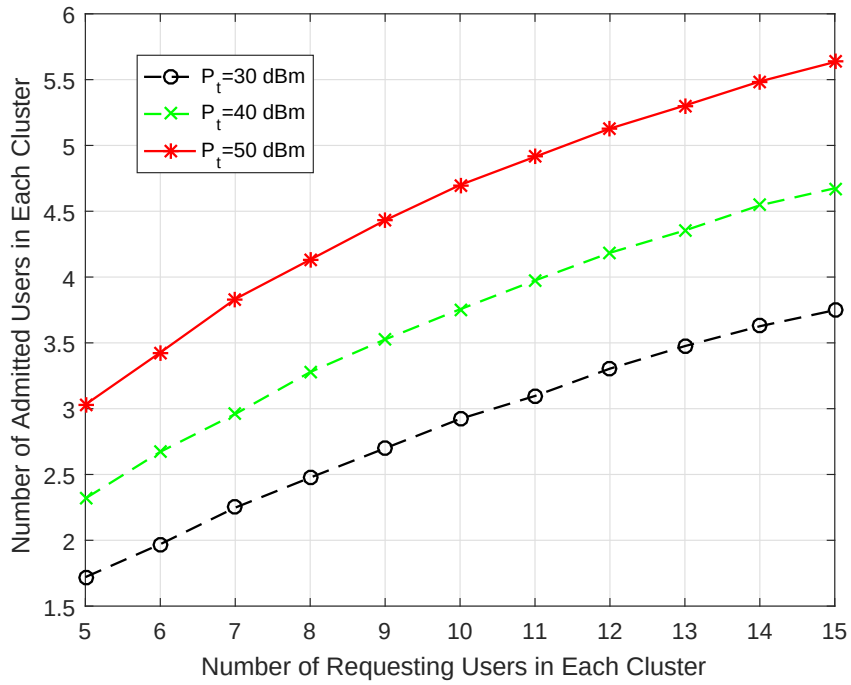


Fig. 2.8: Number of admitted users vs. number of requesting users with different transmit power.

user. Since the total transmit power is fixed, the number of admitted users decreases accordingly. On the other hand, if the total transmit power increases, more users can be admitted, which is verified by the difference in the number of admitted users when the total transmit power is 30 dBm, 40 dBm and 50 dBm, respectively. When the target SINR is 5 dB, about 4 users can be admitted into each cluster even when the total transmit power is 30 dBm, which indicates the effectiveness of the proposed user admission scheme. Further, when the total transmit power is 50 dBm, about 6.5 users on average are admitted to each cluster.

Figs. 2.8 and 2.9 illustrate how the number of admitted users per cluster varies with that of the requesting users per cluster. Specifically, Fig. 2.8 shows results for different transmit powers, while Fig. 2.9 displays results for different target SINRs. Note that the target SINR is set to 10 dB in Fig. 2.8, whereas the total transmit power is set to 35 dBm

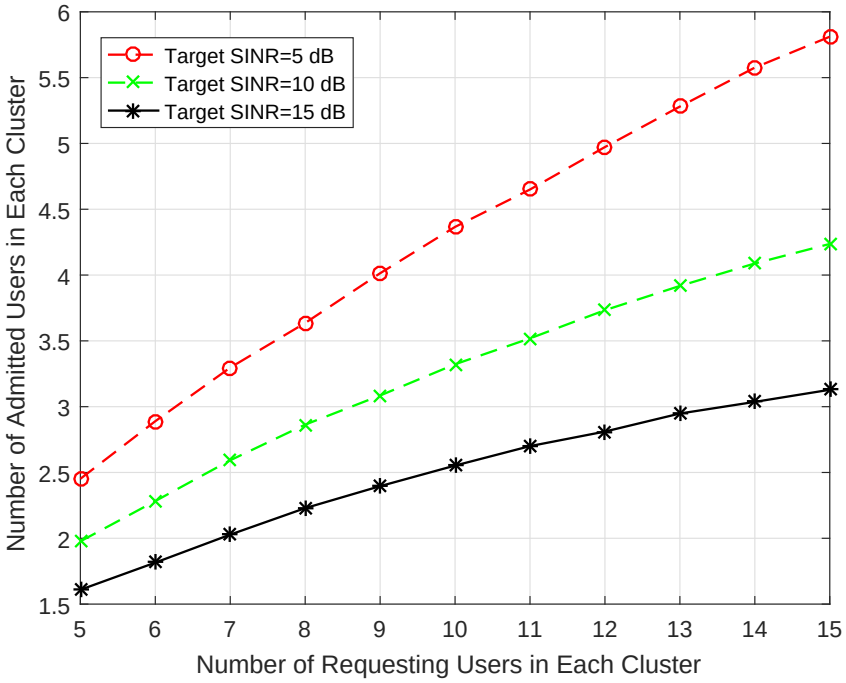


Fig. 2.9: Number of admitted users vs. number of requesting users with different target SINR.

in Fig. 2.9. From these figures, it can be observed that the number of admitted users per cluster grows with that of the requesting users. This is due to the fact that with more users requesting admission, more users are likely to have a better channel. According to the proposed user admission scheme, i.e., (2.29), less power is required to admit one user when it has a good channel gain. Therefore, more users can be admitted with the same total transmit power. Further, as expected, results in Figs. 2.8 and 2.9 show that the number of admitted users per cluster grows with the total transmit power, while it decreases with the target SINR, respectively.

In Figs. 2.10 and 2.11, the performance of the proposed algorithm and exhaustive search is compared. Specifically, the exhaustive search is conducted as follows: first, we consider all possible combinations of the users; then, for each combination, we use (2.29) to allocate the power coefficient to each user, and decide whether this combination

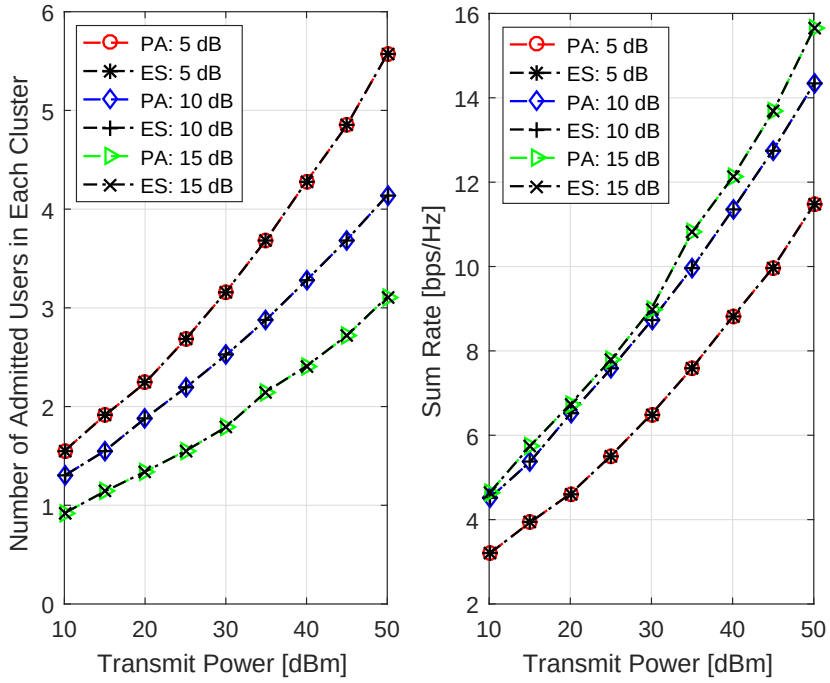


Fig. 2.10: Proposed algorithm vs. exhaustive search when the target SINRs of the users are equal.

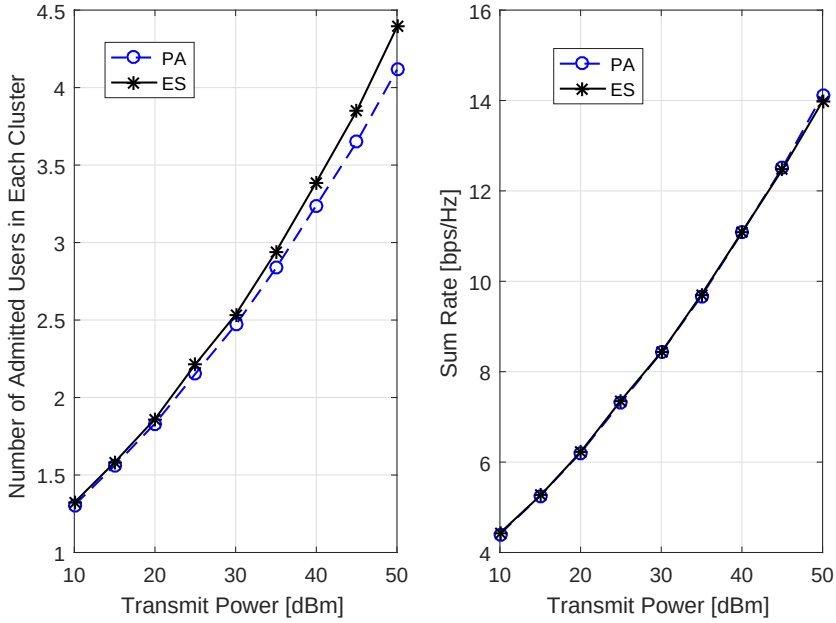


Fig. 2.11: Proposed algorithm vs. exhaustive search when the user target SINRs are different.

is feasible or not; among all feasible combinations, we select the ones with the largest number of users; lastly, from the selected ones, the one with the highest sum rate is chosen. In simulations, the number of requesting users is 8, and results are obtained from 1000 trials. Note that PA and ES in the legend represent the proposed algorithm and exhaustive search, respectively.

In Fig. 2.10, the target SINR of all users is equal, and the number in the legend represents its value. According to Fig. 2.10, it can be seen that the performance of the proposed algorithm is the same as the one of the exhaustive search in terms of both sum rate and number of admitted users for all three target SINRs. In addition, the number of admitted users decreases with the target SINRs, while the sum rate exhibits an opposite trend. The former can be easily explained, whereas the latter is due to the fact that the increase in the data rate of the admitted users dominates the decrease in the number of admitted users.

Furthermore, in Fig. 2.11, the comparison is conducted when the target SINRs are different. Specifically, each user is randomly assigned a target SINR value of 5, 10, or 15 dB. As can be seen from Fig. 2.11, the exhaustive search achieves better result in terms of the number of admitted users per cluster. However, the gap between the proposed algorithm and exhaustive search is minor. In particular, when the transmit power is 50 dBm, the gap reaches a peak, which is only 0.27. On the other hand, as for the sum rate, a similar performance is achieved. Additionally, the complexity of exhaustive search is $N!$, while the proposed algorithm has a low complexity, i.e., linear to the number of users per cluster. To conclude, these results verify the effectiveness of the proposed algorithm also when the users' target SINRs are different.

2.7 Conclusion

This chapter compared the capacity of MIMO-NOMA with that of MIMO-OMA, when multiple users are grouped into a cluster. The superiority of MIMO-NOMA over MIMO-OMA was demonstrated in terms of both sum channel capacity and ergodic sum capacity. Furthermore, the power coefficient value that maximizes the sum rate gap between MIMO-NOMA and MIMO-OMA was derived, when there are two users per cluster. Meanwhile, for two and three users per cluster, numerical results also verified that MIMO-NOMA dominates MIMO-OMA in terms of user fairness. It was also proved that the more users are admitted to the same cluster, the lower is the achieved sum rate, which implies a tradeoff between sum rate and number of admitted users. On this basis, a user admission scheme was proposed, which achieves optimal results in terms of both sum rate and number of admitted users when the SINR thresholds of the users are equal. When the SINR thresholds of the users are different, the proposed scheme still achieves good performance in balancing both criteria. Furthermore, the proposed scheme is of low complexity, i.e., linear in the number of users in each cluster. Finally, the developed analytical results were validated by simulation results.

Appendix

Proof of Lemma 2.1

At the receiver side of user (m, l), the following constraint has to be satisfied in order to implement SIC effectively:

$$R_{m,l}^{k} \geq R_{m,k}^{\mathrm{NOMA}}, \forall k \in \{l+1, \ldots, L\}, \tag{2.31}$$

where $R_{m,l}^{k}$ denotes the data rate of user (m,k) achieved at the receiver (m,l), whereas $R_{m,k}^{\mathrm{NOMA}}$ represents the achievable data rate of user (m,k) at its receiver side. Indeed, the above equation guarantees that user (m,l) can remove the interference of those users with worse channel gains, i.e., $(m,l+1),\ldots,(m,L)$. According to the order of the effective channel gains, i.e., $|\mathbf{v}_{m,l}^{H}\mathbf{H}_{m,l}\mathbf{p}_m|^2 \geq |\mathbf{v}_{m,k}^{H}\mathbf{H}_{m,k}\mathbf{p}_m|^2, \forall k \geq l$, we have

$$\begin{aligned} R_{m,l}^{k} &= \log_2\left(1+\frac{\rho\Omega_{m,k}|\mathbf{v}_{m,l}^{H}\mathbf{H}_{m,l}\mathbf{p}_m|^2}{1+\rho\sum_{i=1}^{k-1}\Omega_{m,i}|\mathbf{v}_{m,l}^{H}\mathbf{H}_{m,l}\mathbf{p}_m|^2}\right) \\ &\geq \log_2\left(1+\frac{\rho\Omega_{m,k}|\mathbf{v}_{m,k}^{H}\mathbf{H}_{m,k}\mathbf{p}_m|^2}{1+\rho\sum_{i=1}^{k-1}\Omega_{m,i}|\mathbf{v}_{m,k}^{H}\mathbf{H}_{m,k}\mathbf{p}_m|^2}\right) \\ &= R_{m,k}^{\mathrm{NOMA}}. \end{aligned} \tag{2.32}$$

Thus, $R_{m,l}^{k} \geq R_{m,k}^{\mathrm{NOMA}}, \forall k \in \{l+1,\ldots,L\}$ is always true. Consequently, the use of SIC is always guaranteed at the receiver (m,l) owing to the ordering of the effective channel gains, and this puts no extra constraints on the system.

Proof of Theorem 2.1

For simplicity of notation, let $K_l = \rho\Omega_{m,l}|\mathbf{v}_{m,l}^{H}\mathbf{H}_{m,l}\mathbf{p}_m|^2$, $l \in \{1,\ldots,L\}$. Theorem 2.1 can be proved via mathematical induction, and the hypothesis is

$$S_{m,L_1}^{\mathrm{OMA}} \leq (\sum_{l=1}^{L_1}\lambda_{m,l})\log_2(1+\frac{\sum_{l=1}^{L_1}K_l}{\sum_{l=1}^{L_1}\lambda_{m,l}}), \tag{2.33}$$

where S_{m,L_1}^{OMA} represents the sum rate for the first L_1 users, $L_1 \in \{1,\ldots,L\}$. Obviously, the first user satisfies the hypothesis, since $S_{m,1}^{\mathrm{OMA}} = R_{m,1}^{\mathrm{OMA}} = \lambda_{m,1}\log_2(1+\frac{K_1}{\lambda_{m,1}})$.

Then, let us consider the case of $L_2 = L_1 + 1$, and we have

$$
\begin{aligned}
&S_{m,L_2}^{\mathrm{OMA}} \\
&= S_{m,L_1}^{\mathrm{OMA}} + \lambda_{m,L_2} \log_2(1 + \frac{K_{L_2}}{\lambda_{m,L_2}}) \\
&\leq (\sum_{l=1}^{L_1} \lambda_{m,l}) \log_2(1 + \frac{\sum_{l=1}^{L_1} K_l}{\sum_{l=1}^{L_1} \lambda_{m,l}}) + \lambda_{m,L_2} \log_2(1 + \frac{K_{L_2}}{\lambda_{m,L_2}}) \\
&= (\sum_{l=1}^{L_2} \lambda_{m,l}) \Bigg[\frac{\sum_{l=1}^{L_1} \lambda_{m,l}}{\sum_{l=1}^{L_2} \lambda_{m,l}} \log_2(1 + \frac{\sum_{l=1}^{L_1} K_l}{\sum_{l=1}^{L_2} \lambda_{m,l}} \frac{\sum_{l=1}^{L_2} \lambda_{m,l}}{\sum_{l=1}^{L_1} \lambda_{m,l}}) \\
&+ \frac{\lambda_{m,L_2}}{(\sum_{l=1}^{L_2} \lambda_{m,l})} \log_2(1 + \frac{K_{L_2}}{\sum_{l=1}^{L_2} \lambda_{m,l}} \frac{\sum_{l=1}^{L_2} \lambda_{m,l}}{\lambda_{m,L_2}}) \Bigg].
\end{aligned}
\tag{2.34}
$$

Let $\lambda = \frac{\sum_{l=1}^{L_1} \lambda_{m,l}}{\sum_{l=1}^{L_2} \lambda_{m,l}}$, then $1 - \lambda = \frac{\lambda_{m,L_2}}{(\sum_{l=1}^{L_2} \lambda_{m,l})}$. In addition, let $K_1' = \frac{\sum_{l=1}^{L_1} K_l}{\sum_{l=1}^{L_2} \lambda_{m,l}}$ and $K_2' = \frac{K_{L_2}}{\sum_{l=1}^{L_2} \lambda_{m,l}}$. The polynomial in the bracket can be reformulated as $\lambda \log_2(1 + \frac{K_1'}{\lambda}) + (1 - \lambda) \log_2(1 + \frac{K_2'}{1-\lambda})$, which has the same form as [15, eq. (12)]. According to Lemma 2.2, it can be written as $\log_2(1 + \frac{\sum_{l=1}^{L_2} K_l}{\sum_{l=1}^{L_2} \lambda_{m,l}})$, satisfying $\frac{\sum_{l=1}^{L_1} K_l}{\sum_{l=1}^{L_1} \lambda_{m,l}} = \frac{K_{L_2}}{\lambda_{m,L_2}}$. Thus, we have $S_{m,L_2}^{\mathrm{OMA}} \leq (\sum_{l=1}^{L_2} \lambda_{m,l}) \log_2(1 + \frac{\sum_{l=1}^{L_2} K_l}{\sum_{l=1}^{L_2} \lambda_{m,l}})$, which also fits the hypothesis.

Lastly, we consider the case for L users. Due to $\sum_{l=1}^{L} \lambda_{m,l} = 1$, we have $S_{m,L}^{\mathrm{OMA}} \leq \log_2(1 + \sum_{l=1}^{L} K_l) = \log_2(1 + \sum_{l=1}^{L} \rho \Omega_{m,l} |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2)$. Here Theorem 2.1 is proved. Moreover, it is easy to conclude that the equality is achieved when $\frac{\Omega_{m,1} |\mathbf{v}_{m,1}^H \mathbf{H}_{m,1} \mathbf{p}_m|^2}{\lambda_{m,1}} = \cdots = \frac{\Omega_{m,l} |\mathbf{v}_{m,L}^H \mathbf{H}_{m,L} \mathbf{p}_m|^2}{\lambda_{m,L}}$. Correspondingly, we have $\lambda_{m,l} = \frac{\Omega_{m,l} |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2}{\sum_1^L \Omega_{m,l} |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2}, \forall\, l \in \{1, \ldots, L\}$.

Proof of Lemma 2.3

According to inequality 2.6, we have $\rho \Omega_{m,k} |\mathbf{v}_{m,k}^H \mathbf{H}_{m,k} \mathbf{p}_m|^2 \geq \rho \Omega_{m,k} |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2, \forall k \leq l$. Consequently, it can be concluded that

$$
\frac{1 + \rho \sum_{k=1}^{l} \Omega_{m,k} |\mathbf{v}_{m,k}^H \mathbf{H}_{m,k} \mathbf{p}_m|^2}{1 + \rho \sum_{k=1}^{l} \Omega_{m,k} |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2} \geq 1, l \in \{1, \ldots, L\}. \tag{2.35}
$$

Further, the above equation can be used to obtain the lower bound for the sum rate for MIMO-NOMA via mathematical induction, and the hypothesis is that the sum rate for the first l users, denoted as $S_{m,l}^{\text{NOMA}}$ is bounded by

$$S_{m,l}^{\text{NOMA}} \geq \log_2(1+\rho\sum_{k=1}^{l}\Omega_{m,k}|\mathbf{v}_{m,k}^H\mathbf{H}_{m,k}\mathbf{p}_m|^2). \tag{2.36}$$

Clearly, the first user satisfies (2.36), since $S_{m,1}^{\text{NOMA}} = R_{m,1}^{\text{NOMA}} = \log_2(1+\rho\Omega_{m,1}|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2) \geq \log_2(1+\rho\Omega_{m,1}|\mathbf{v}_{m,1}^H\mathbf{H}_{m,1}\mathbf{p}_m|^2)$.

Next, the case for $l+1$ users is proved as follows:

$$\begin{aligned}
S_{m,l+1}^{\text{NOMA}} &= S_{m,l}^{\text{NOMA}} + R_{m,l+1}^{\text{NOMA}} \\
&\geq \log_2(1+\rho\sum_{k=1}^{l}\Omega_{m,k}|\mathbf{v}_{m,k}^H\mathbf{H}_{m,k}\mathbf{p}_m|^2) \\
&+ \log_2(1+\frac{\rho\Omega_{m,l+1}|\mathbf{v}_{m,l+1}^H\mathbf{H}_{m,l+1}\mathbf{p}_m|^2}{1+\rho\sum_{k=1}^{l}\Omega_{m,k}|\mathbf{v}_{m,l+1}^H\mathbf{H}_{m,l+1}\mathbf{p}_m|^2}) \\
&= \log_2(1+\rho\sum_{k=1}^{l}\Omega_{m,k}|\mathbf{v}_{m,k}^H\mathbf{H}_{m,k}\mathbf{p}_m|^2 \\
&+ \rho\Omega_{m,l+1}|\mathbf{v}_{m,l+1}^H\mathbf{H}_{m,l+1}\mathbf{p}_m|^2 \\
&\times \frac{(1+\rho\sum_{k=1}^{l}\Omega_{m,k}|\mathbf{v}_{m,k}^H\mathbf{H}_{m,k}\mathbf{p}_m|^2)}{1+\rho\sum_{k=1}^{l}\Omega_{m,k}|\mathbf{v}_{m,l+1}^H\mathbf{H}_{m,l+1}\mathbf{p}_m|^2}) \\
&\geq \log_2(1+\rho\sum_{k=1}^{l+1}\Omega_{m,k}|\mathbf{v}_{m,k}^H\mathbf{H}_{m,k}\mathbf{p}_m|^2),
\end{aligned} \tag{2.37}$$

where the last inequality comes from (2.35).

Thus, when all L users are considered, we have $S_{m,L}^{\text{NOMA}} \geq \log_2(1+\rho\sum_{k=1}^{L}\Omega_{m,k}|\mathbf{v}_{m,k}^H\mathbf{H}_{m,k}\mathbf{p}_m|^2)$. Hence, Lemma 2.3 is proved.

Proof of Theorem 2.3

Consider the case in which only l users can be admitted to the mth cluster when employing the proposed user admission scheme. Suppose there exists an alternate scheme, which also admits l users, denoted as $a_1, a_2, \ldots, a_l$. Theorem 2.3 can be proved through contradiction.

Specifically, the proof consists of two steps: 1) it is shown that the sum power required by the alternate scheme always exceeds that of the proposed scheme; and 2) based on (1), assume that the alternate scheme can admit an extra user, this user should also be admitted by the proposed scheme, which conflicts with the proposition that only l users can be admitted by the proposed scheme. Consequently, no other scheme can admit a larger number of users than the proposed one.

Step 1: The power coefficients of the proposed scheme and the alternate one are denoted as $\Omega_1, \Omega_2, \ldots, \Omega_l$, and $\Omega_{a_1}, \Omega_{a_2}, \ldots, \Omega_{a_l}$, respectively. For notational simplicity, let $G_k = |\mathbf{v}_k^H \mathbf{H}_k \mathbf{p}|^2, k \in \{1\ , 2\ , \ldots, l\}$, and $G_{a_k} = |\mathbf{v}_{a_k}^H \mathbf{H}_{a_k} \mathbf{p}|^2, a_k \in \{a_1, a_2, \ldots, a_l\}$. Without loss of generality, the admitted l users for the alternate scheme are also ranked in a descending order according to their channel gains, i.e., $G_{a_1} \geq \cdots \geq G_{a_l}$. Thus, it can be easily observed that $G_{a_k} \leq G_k$, since $k \leq a_k$ due to the channel order and user admission order of both schemes. Moreover, according to (2.26) and (2.29), we have $\Omega_{a_k} \geq \Gamma_{a_k} \sum_{i=1}^{k-1} \Omega_{a_i} + \frac{\Gamma_{a_k}}{\rho G_{a_k}}$, and $\Omega_k = \Gamma_k \sum_{i=1}^{k-1} \Omega_i + \frac{\Gamma_k}{\rho G_k}$, respectively. After some algebraic manipulations, the sums of the power coefficients for the proposed scheme and the alternate one can be expressed as

$$\Psi = \sum_{k=1}^{l} \frac{\Gamma k}{\rho G_k} \prod_{i=k+1}^{l} (\Gamma_i + 1) \tag{2.38a}$$

$$\Psi_a \geq \sum_{a_k=1}^{a_l} \frac{\Gamma a_k}{\rho G_{a_k}} \prod_{a_i=a_k+1}^{a_l} (\Gamma_{a_i} + 1), \tag{2.38b}$$

where Ψ and Ψ_a denote the sums of the power coefficients for the proposed scheme and

the alternate one, respectively.

By using (2.30a), (2.30b) and $G_{a_k} \leq G_k$, it can be easily obtained that $\frac{\Gamma k}{G_k} \leq \frac{\Gamma a_k}{G_{a_k}}$, and $\prod_{i=k+1}^{l}(\Gamma_i + 1) \leq \prod_{a_i=a_k+1}^{a_l}(\Gamma_{a_i} + 1)$. Thus, $\Psi \leq \Psi_a$, which means that to admit the same number of users, the proposed scheme requires the minimum power.

Step 2: Suppose the alternate scheme can admit an extra user, a_{l+1}, whose power coefficient and channel gain are denoted as $\Omega_{a_{l+1}}$ and $G_{a_{l+1}}$, respectively. According to (2.26) and (2.24b), we have $\Omega_{a_{l+1}} \geq \Gamma_{a_{l+1}}\Psi_a + \frac{\Gamma_{a_{l+1}}}{\rho G_{a_{l+1}}}$, which must satisfy $\Omega_{a_{l+1}} + \Psi_a \leq 1$. On this basis, it is easy to verify that user a_{l+1} can also be admitted by the proposed scheme, since $\Omega'_{a_{l+1}} + \Psi \leq \Omega_{a_{l+1}} + \Psi_a \leq 1$, where $\Omega'_{a_{l+1}} = \Gamma_{a_{l+1}}\Psi + \frac{\Gamma_{a_{l+1}}}{\rho G_{a_{l+1}}}$ denotes the power coefficient of user a_{l+1} under the proposed scheme. Clearly, this conflicts with the proposition that only l users can be admitted by the proposed scheme.

References

[1] 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 Veh. Technol. Conf.*, Dresden, Germany, Jun. 2013, pp. 1–5.

[2] L. Dai, B. Wang, Y. Yuan, S. Han, C. l. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[3] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surv. Tuts.*, vol. pp, no. 99, pp. 1–1, Oct. 2016.

[4] M. Zeng, G. I. Tsiropoulos, O. A. Dobre, and M. H. Ahmed, "Power allocation for cognitive radio networks employing non-orthogonal multiple access," in *Proc. IEEE Global Telecommun. Conf.*, Washington DC, USA, Dec. 2016.

[5] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5G wireless networks," in *Proc. Int. Symp. Wireless Commun. Systems*, Barcelona, Spain, Aug. 2014, pp. 781–785.

[6] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag. – Feature Topic on LTE Evolution,* to appear. *arXiv preprint arXiv:1511.08610*, 2015.

[7] S. M. R. Islam, M. Zeng, and O. A. Dobre, "Noma in 5G systems: Exciting possibilities for enhancing spectral efficiency," *IEEE 5G Tech. Focus,* to appear, 2017.

[8] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[9] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication.* Cambridge, UK: Cambridge University Press, 2005.

[10] 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, Dec. 2014.

[11] Q. Sun, S. Han, I. Chin-Lin, and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 4, pp. 405–408, Dec. 2015.

[12] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[13] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, Jun. 2016.

[14] Y. Liu, G. Pan, H. Zhang, and M. Song, "On the capacity comparison between MIMO-NOMA and MIMO-OMA," *IEEE Access*, vol. 4, no. 6, pp. 2123–2129, Jul. 2016.

[15] M. Zeng, Y. Animesh, 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.,* DOI: 10.1109/LWC.2017.2712149.

[16] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7244–7257, Nov. 2016.

[17] Z. Ding and H. V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 629–633, May 2016.

Chapter 3

Energy-Efficient Power Allocation for MIMO-NOMA with Multiple Users in a Cluster

3.1 Abstract

In this chapter, energy-efficient PA is investigated for a MIMO-NOMA system with multiple users in a cluster. To ensure the QoS for the users, a minimum rate requirement is pre-defined for each user. Because of the QoS requirement, it is first necessary to determine whether the considered energy efficiency (EE) maximization problem is feasible or not, by comparing the total transmit power with the required power for satisfying the QoS of the users. If feasible, a closed-form solution is provided for the corresponding sum rate maximization problem, and on this basis, the EE maximization problem is solved by applying non-convex fractional programming. Otherwise, a low complexity user admission scheme is proposed, which admits users one by one following the ascending order of the required power for satisfying the QoS. Numerical results are presented to validate the

effectiveness of the proposed energy-efficient PA strategy and user admission scheme.

3.2 Introduction

Power-domain NOMA has been widely considered as a promising candidate for the next generation of wireless communication systems [1–5]. By applying superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver, NOMA multiplexes multiple users in the power domain, to access the same time-frequency resource. When compared with conventional OMA scheme, NOMA achieves higher spectral efficiency (SE) [6–8]. The authors in [6] show via simulation that NOMA provides a larger sum rate than OMA, while in [7], the authors prove the dominance of NOMA over OMA by comparing their achievable rate regions. Furthermore, the authors in [8] validate that NOMA achieves higher ergodic sum rate than OMA for a cellular downlink scenario with randomly deployed users.

However, the above works only consider SISO channels. Recently, MIMO has also been integrated into NOMA to further enhance the SE [9–13]. For MIMO-NOMA systems, users are usually paired into clusters to reduce the complexity of SIC at the receiver, with users in the same cluster sharing a common beamformer. The authors in [9] and [10] show that MIMO-NOMA achieves larger sum rates than MIMO-OMA for a two-user multi-cluster system, while [11] and [12] further validate that this performance advantage still holds for a multi-user per cluster system. Note that the above works only consider power allocation (PA) within each cluster, by allocating equal power to each cluster. In [13], the authors propose a beamspace MIMO-NOMA scheme for a millimeter wave system, which allows power to be distributed among clusters. Simulation results illustrate that the proposed beamspace MIMO-NOMA achieves higher SE when compared with existing beamspace MIMO-OMA. In [14], the authors extend the study of MIMO-NOMA from

single cell to multicell and investigate the precoder design. Numerical results show that the proposed NOMA design improves both edge and sum throughput compared with conventional OMA.

Nevertheless, the studies above mainly focus on the SE of NOMA systems. As energy efficiency (EE) becomes one of the major concerns for 5G, it is of interest to investigate the EE for NOMA [15–17]. In [15], the authors study the joint subchannel assignment and PA to maximize the EE for a multi-carrier NOMA system. The obtained simulation results show that NOMA achieves higher SE and EE than OMA. However, this work is only applicable to systems with two users per cluster. In [16], EE is studied under a single-carrier multi-user NOMA system with QoS requirement for each user. A PA algorithm is proposed based on non-convex fractional programming, and numerical results validate that NOMA exhibits better EE performance than OMA. Note that both [15] and [16] consider SISO systems. Since current and future communication systems rely on the multiple antenna (MIMO) structure, the EE under MIMO-NOMA is of interest. In [17], the authors investigate the EE in a millimeter wave massive MIMO-NOMA system with a low-complexity radio frequency (RF) chain structure at the BS. A hybrid analog/digital precoding scheme is proposed first. Based on this, a PA problem aiming to maximize the EE is formulated under users' quality of service (QoS) requirements and per-cluster equal power constraint, and an iterative algorithm is proposed to obtain an optimal PA.

To the best of our knowledge, none of the existing works has studied the EE for a multi-cluster MIMO-NOMA with multiple users per cluster. Moreover, most existing studies assume that the total transmit power is large enough to satisfy the QoS requirements for all users, without considering the situation when this assumption does not hold [15–17]. Toward filling this research gap, the contributions of this chapter are summarized as follows:

- This chapter studies the EE for a multi-cluster multi-user MIMO-NOMA system

with pre-defined QoS for each user in a systematic way: it is first determined whether all users can be admitted or not by comparing the total transmit power with the power required to satisfy the QoS for all users; when all users can be admitted, the objective is to maximize the EE of the system; otherwise, the objective is to maximize the number of admitted users;

- For the EE maximization problem, global PA is considered: it is first determined how to allocate power within each cluster; on this basis, the relationship between the sum rate increment and required extra power for each cluster is derived; by exploiting this relationship, a water-filling-like optimal PA is proposed to maximize the sum rate of the system under any given total power; lastly, it is proved that the EE function is pseudo-concave over the final "water" level, and can be solved accordingly;

- For the user admission problem, a low complexity algorithm is proposed, which admits the users one by one following the ascending order of the required power to satisfy their QoS; further analysis on its optimality and complexity is provided, which validates the effectiveness of the proposed algorithm.

The rest of the chapter is organized as follows. The system model and problem formulation are introduced in Section 3.2. The proposed energy-efficient PA strategy and user admission scheme are elaborated in Section 3.3. Simulation results are shown in Section 3.4, while conclusions are finally drawn in Section 3.5.

3.3 System Model and Problem Formulation

3.3.1 System Model

A downlink multi-user MIMO system is considered in this chapter, in which the BS equipped with M antennas sends data to multiple receivers, each equipped with N antennas. The total number of users in the system is $M \times L$, which are grouped into M clusters randomly with $L\,(L \geq 2)$ users per cluster. NOMA is applied among the users in the same cluster. The channel matrix between the BS and the lth user in the mth cluster, i.e., user (m, l) $(m \in \{1, \ldots, M\}, l \in \{1, \ldots, L\})$ is denoted as $\mathbf{H}_{m,l} \in \mathbb{C}^{N \times M}$, which is assumed to be quasi-static independent and identically distributed (i.i.d.). In addition, the precoding matrix used by the BS is denoted as $\mathbf{P} \in \mathbb{C}^{M \times M}$, whereas the detection vector for user (m, l) is represented by $\mathbf{v}_{m,l} \in \mathbb{C}^{N \times 1}$. They should satisfy: a) $\mathbf{P} = \mathbf{I}_M$, with $\mathbf{I}_M$ denoting the $M \times M$ identity matrix; b) $|\mathbf{v}_{m,l}|^2 = 1$ and $\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_k = 0$ for any $k \neq m$, where $\mathbf{p}_k$ is the kth column of $\mathbf{P}$ [11]. Note that the number of antennas should satisfy $N \geq M$ to make this feasible. Because of the zero-forcing (ZF) based detection design, the inter-cluster interference can be removed even when there exist multiple users in a cluster. Note that only a scalar value $|\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2$ needs to be fed back to the BS from user (m, l).

For the considered MIMO-NOMA scheme, the BS multiplexes the intended signals for all users at the same frequency and time resource. Therefore, the corresponding transmitted signals from the BS can be expressed as

$$\mathbf{x} = \mathbf{P}\mathbf{s}, \tag{3.1}$$

where the information-bearing vector $\mathbf{s} \in \mathbb{C}^{M\times 1}$ can be further written as

$$\mathbf{s} = \begin{bmatrix} \sqrt{P_{\max}\Omega_{1,1}}s_{1,1} + \cdots + \sqrt{P_{\max}\Omega_{1,L}}s_{1,L} \\ \vdots \\ \sqrt{P_{\max}\Omega_{M,1}}s_{M,1} + \cdots + \sqrt{P_{\max}\Omega_{M,L}}s_{M,L} \end{bmatrix}, \tag{3.2}$$

where $s_{m,l}$ and $\Omega_{m,l}$ denote the signal and power allocation coefficient for user (m,l), satisfying $\sum_{m=1}^{M}\sum_{l=1}^{L}\Omega_{m,l} \leq 1$. $P_{\max}$ denotes the total transmit power for the BS.

Accordingly, at user (m,l), the observed signal is given by

$$\mathbf{y}_{m,l} = \mathbf{H}_{m,l}\mathbf{P}\mathbf{s} + \mathbf{n}_{m,l}, \tag{3.3}$$

where $\mathbf{n}_{m,l}$ is the independent and identically distributed (i.i.d.) additive white Gaussian (AWGN) noise vector, $\mathcal{CN}(0,\sigma^2\mathbf{I})$.

By applying the detection vector $\mathbf{v}_{m,l}$ on the observed signal, (3.3) can be expressed as

$$\mathbf{v}_{m,l}^H\mathbf{y}_{m,l} = \mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_m\sum_{l=1}^{L}\sqrt{P_{\max}\Omega_{m,l}}s_{m,l} + \underbrace{\sum_{k=1,k\neq m}^{M}\mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_k\mathbf{s}_k}_{\text{interference from other clusters}} + \mathbf{v}_{m,l}^H\mathbf{n}_{m,l}, \tag{3.4}$$

where $\mathbf{s}_k$ denotes the kth row of $\mathbf{s}$.

Owing to the constraint[1] on the detection vector, i.e., $\mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_k = 0$ for any $k \neq m$, (3.4) can be simplified as

$$\mathbf{v}_{m,l}^H\mathbf{y}_{m,l} = \mathbf{v}_{m,l}^H\mathbf{H}_{m,l}\mathbf{p}_m\sum_{l=1}^{L}\sqrt{P_{\max}\Omega_{m,l}}s_{m,l} + \mathbf{v}_{m,l}^H\mathbf{n}_{m,l}. \tag{3.5}$$

[1]Due to the specific selection of $\mathbf{P}$, this constraint is further reduced to $\mathbf{v}_{m,l}^H\tilde{\mathbf{H}}_{m,l} = 0$, where $\tilde{\mathbf{H}}_{m,l} = [\mathbf{h}_{1,ml}\cdots\mathbf{h}_{m-1,ml}\,\mathbf{h}_{m+1,ml}\cdots\mathbf{h}_{M,ml}]$ and $\mathbf{h}_{i,ml}$ is the ith column of $\mathbf{H}_{m,l}$ [11]. Hence, $\mathbf{v}_{m,l}$ can be expressed as $\mathbf{U}_{m,l}\mathbf{w}_{m,l}$, where $\mathbf{U}_{m,l}$ is the matrix consisting of the left singular vectors of $\tilde{\mathbf{H}}_{m,l}$ corresponding to the non-zero singular values, and $\mathbf{w}_{m,l}$ is the maximum ratio combining vector expressed as $\mathbf{U}_{m,l}^H\mathbf{h}_{m,ml}/|\mathbf{U}_{m,l}^H\mathbf{h}_{m,ml}|$.

Without loss of generality, the effective channel gains are ordered as [11]

$$|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1} \mathbf{p}_m|^2 \geq \cdots \geq |\mathbf{v}_{m,L}^H \mathbf{H}_{m,L} \mathbf{p}_m|^2. \tag{3.6}$$

At the receiver, each user conducts SIC to remove the interference from the users with worse channel gains, i.e., the interfence from user $(m, l+1), \ldots, (m, L)$ is removed by user (m, l).[2] As a result, the achieved data rate at user (m, l) is given by [11]

$$R_{m,l} = \log_2 \left(1 + \frac{\rho \Omega_{m,l} |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2}{1 + \rho \sum_{k=1}^{l-1} \Omega_{m,k} |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2} \right), \tag{3.7}$$

where $\rho = P_{\max}/\sigma^2$ denotes the transmit signal-to-noise ratio (SNR).

3.3.2 Problem Formulation

The total power consumption is comprised of two parts: the fixed circuit power consumption P_{c}, and the flexible transmit power $P_{\mathrm{t}} = P_{\max} \sum_{m=1}^{M} \sum_{l=1}^{L} \Omega_{m,l}$. Similar to [16], we define the EE of the system as

$$\eta_{\mathrm{EE}} = \frac{R^{\mathrm{sum}}}{P_{\mathrm{c}} + P_{\mathrm{t}}}, \tag{3.8}$$

where $R^{\mathrm{sum}} = \sum_{m=1}^{M} \sum_{l=1}^{L} R_{m,l}$ denotes the achievable sum rate.

We aim to maximize the EE of the system when each user has a pre-defined minimum rate. The considered problem can be formulated as:

$$\max_{\Omega_{m,l}} \ \eta_{\mathrm{EE}} \tag{3.9a}$$

$$\text{s.t.} \ \ R_{m,l} \geq R_{m,l}^{\min}, m \in \{1, \cdots, M\}, l \in \{1, \cdots, L\} \tag{3.9b}$$

$$\sum_{m=1}^{M} \sum_{l=1}^{L} \Omega_{m,l} \leq 1, \tag{3.9c}$$

[2] [12] proves that SIC is guaranteed to be successful.

where (3.9b) and (3.9c) represent the users' minimum rate requirements and the transmit power constraint, respectively.

3.4 Proposed Solution

Owing to the existence of the minimum rate requirements, i.e., (3.9b), the formulated problem (3.9) may be infeasible when the transmit power is not large enough. In this case, instead of EE maximization, maximizing the number of admitted users makes more sense. As such, it is of importance to determine the feasibility of problem (3.9), which can be done by comparing the total transmit power constraint with the minimum power required to satisfy the minimum rate requirements of all users. The minimum required power can be expressed as

$$P_{\text{req}} = P_{\max} \sum_{m=1}^{M} \sum_{l=1}^{L} \Omega_{m,l}^{\min}, \tag{3.10}$$

where $\Omega_{m,l}^{\min} = (2^{R_{m,l}^{\min}} - 1)(\sum_{k=1}^{l-1} \Omega_{m,k}^{\min} + \frac{1}{\rho |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2})$ is the minimum required power to satisfy the QoS requirement of user (m, l) [18, (14)]. As a result, if

$$P_{\text{req}} \leq P_{\max} \Longleftrightarrow \sum_{m=1}^{M} \sum_{l=1}^{L} \Omega_{m,l}^{\min} \leq 1, \tag{3.11}$$

problem (3.9) is feasible and vice versa.

3.4.1 EE Maximization when (3.9) is Feasible

The objective function in (3.9) is of fractional form; hence (3.9) is a non-convex problem and obtaining an optimal solution is non-trival. To solve it in a tractable way, we first turn to the corresponding SE maximization problem. According to the definition of EE in (3.8), to maximize the EE, we need to maximize the corresponding SE under any given power of $P_{\text{f}}, P_{\text{f}} \in [P_{\text{req}}, P_{\max}]$, and then select the appropriate value of P_{f}.

The SE maximization problem can be formulated as

$$\max_{\Omega_{m,l}} \; R^{\text{sum}} \tag{3.12a}$$

$$\text{s.t.} \;\; R_{m,l} \geq R_{m,l}^{\min}, m \in \{1, \cdots, M\}, \tag{3.12b}$$

$$l \in \{1, \cdots, L\}$$

$$P_{\max} \sum_{m=1}^{M} \sum_{l=1}^{L} \Omega_{m,l} \leq P_{\text{f}}. \tag{3.12c}$$

Note that problem (3.12) is still non-convex due to the non-concavity involved in the objective function. In order to proceed towards an optimal solution, we first determine the PA within each cluster and then across clusters. For PA within each cluster, the following lemma provides some insight:

Lemma 3.1. *Under any given total power constraint for a cluster,*[3] *in order to maximize the cluster sum rate, PA in the cluster should be conducted such that each user (except the first user) receives the amount of power such that its QoS requirement is just satisfied, while the first user receives the remaining power.*

Proof: To prove the lemma, we first prove that transferring power from any other user to the first user leads to an increased sum rate. Assume that the power transfer happens between the nth user and the 1st user, and denote the extra power coefficient as ΔP_{tr}. According to (3.7), the rates of the users with worse channel gains than the nth user remain unchanged, since the total interference does not change. Thus, when comparing the two cluster sum rates, we only need to compare the first n users.

[3]The total power is large enough to satisfy the QoS requirements of all users in the cluster.

The sum rate of the first n users before power transfer can be expressed as

$$\begin{aligned}\sum_{l=1}^{n} R_{m,l} &= \sum_{l=1}^{n} \log_2 \left(\frac{1+\rho \sum_{k=1}^{l} \Omega_{m,k} |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2}{1+\rho \sum_{k=1}^{l-1} \Omega_{m,k} |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2} \right) \\ &= \log_2 \Bigg(\prod_{l=1}^{n-1} \frac{1+\rho \sum_{k=1}^{l} \Omega_{m,k} |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2}{1+\rho \sum_{k=1}^{l} \Omega_{m,k} |\mathbf{v}_{m,l+1}^H \mathbf{H}_{m,l+1} \mathbf{p}_m|^2} \\ &\quad \times (1+\rho \sum_{k=1}^{n} \Omega_{m,k} |\mathbf{v}_{m,n}^H \mathbf{H}_{m,n} \mathbf{p}_m|^2) \Bigg).\end{aligned}$$

Likewise, the sum rate of the first n users after power transfer can be expressed as

$$\begin{aligned}\sum_{l=1}^{n} R'_{m,l} = \log_2 \Bigg(& \prod_{l=1}^{n-1} \frac{1+\rho(\Delta P_{\mathrm{tr}} + \sum_{k=1}^{l} \Omega_{m,k}) |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2}{1+\rho(\Delta P_{\mathrm{tr}} + \sum_{k=1}^{l} \Omega_{m,k}) |\mathbf{v}_{m,l+1}^H \mathbf{H}_{m,l+1} \mathbf{p}_m|^2} \\ & \times (1+\rho \sum_{k=1}^{n} \Omega_{m,k} |\mathbf{v}_{m,n}^H \mathbf{H}_{m,n} \mathbf{p}_m|^2) \Bigg).\end{aligned} \tag{3.13}$$

Since $|\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2 \geq |\mathbf{v}_{m,l+1}^H \mathbf{H}_{m,l+1} \mathbf{p}_m|^2$, it can be easily verified that

$$\begin{aligned}&\frac{1+\rho \sum_{k=1}^{l} \Omega_{m,k} |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2}{1+\rho \sum_{k=1}^{l} \Omega_{m,k} |\mathbf{v}_{m,l+1}^H \mathbf{H}_{m,l+1} \mathbf{p}_m|^2} \\ &< \frac{1+\rho(\Delta P_{\mathrm{tr}} + \sum_{k=1}^{l} \Omega_{m,k}) |\mathbf{v}_{m,l}^H \mathbf{H}_{m,l} \mathbf{p}_m|^2}{1+\rho(\Delta P_{\mathrm{tr}} + \sum_{k=1}^{l} \Omega_{m,k}) |\mathbf{v}_{m,l+1}^H \mathbf{H}_{m,l+1} \mathbf{p}_m|^2}.\end{aligned} \tag{3.14}$$

Therefore, we can prove that $\sum_{l=1}^{n} R_{m,l} < \sum_{l=1}^{n} R'_{m,l}$, which demonstrates that transferring power from other users to the first user yields a larger sum rate. On the other hand, each user should also satisfy its QoS constraint. Combining these two facts, we can conclude that Lemma 3.1 holds. ■

Remark. *The above analysis can be extended to show that transferring power from any user to another user with better channel gains leads to an increased sum rate. This implies that the users with better channel gains have a higher priority than their counterparts. In the user admission section, this property of NOMA is further exploited.*

The above lemma shows how to allocate power within a cluster. Now we consider

PA across clusters. We first allocate the power such that each user's QoS requirement is just satisfied, which requires the power of P_{req}. Correspondingly, the remaining power is denoted as $P_{\text{rem}} = P_{\text{f}} - P_{\text{req}}$. Then, we allocate the remaining power across clusters to maximize the system sum rate. To determine how to allocate power across clusters, the intuition is to compare how much additional power is needed for each cluster when increasing its sum rate by the same unit. The following lemma provides the details:

Lemma 3.2. *Denote the achieved rate of user (m, l) as $\hat{R}_{m,l}$ ($\hat{R}_{m,l} \geq R_{m,l}^{min}$), the additional power required for increasing the sum rate of the mth cluster by ΔR is given by*

$$\Delta P_m = (2^{\Delta R} - 1)\frac{P_{\max} 2^{\sum_{l=1}^{L} \hat{R}_{m,l}}}{\rho |\mathbf{v}_{m,1}^H \mathbf{H}_{m,1} \mathbf{p}_m|^2}. \tag{3.15}$$

Proof: We prove Lemma 3.2 by mathematical induction. Starting with two users per cluster, according to (3.7), we have the following:

$$\hat{\Omega}_{m,1} = \frac{2^{\hat{R}_{m,1}} - 1}{\rho |\mathbf{v}_{m,1}^H \mathbf{H}_{m,1} \mathbf{p}_m|^2} \tag{3.16a}$$

$$\begin{aligned} \hat{\Omega}_{m,2} &= \frac{(2^{\hat{R}_{m,2}} - 1)(1 + \rho \Omega_{m,1} |\mathbf{v}_{m,2}^H \mathbf{H}_{m,2} \mathbf{p}_m|^2)}{\rho |\mathbf{v}_{m,2}^H \mathbf{H}_{m,2} \mathbf{p}_m|^2} \\ &= \frac{2^{\hat{R}_{m,2}} - 1}{\rho |\mathbf{v}_{m,2}^H \mathbf{H}_{m,2} \mathbf{p}_m|^2} + \frac{(2^{\hat{R}_{m,1}} - 1)(2^{\hat{R}_{m,2}} - 1)}{\rho |\mathbf{v}_{m,1}^H \mathbf{H}_{m,1} \mathbf{p}_m|^2}. \end{aligned} \tag{3.16b}$$

According to Lemma 3.1, when some additional power is added to the mth cluster, only the rate of the first user will change, while others remain fixed. Thus, when there is ΔR sum rate increment for the mth cluster, it is only added to $R_{m,1}$, resulting in the change from $\hat{R}_{m,1}$ to $\hat{R}_{m,1} + \Delta R$. Update $R_{m,1}$ in (3.16), and after some algebraic manipulations, the additional power required is given by

$$\Delta P_m^{(2)} = (2^{\Delta R} - 1)\frac{P_{\max} 2^{\sum_{l=1}^{2} \hat{R}_{m,l}}}{\rho |\mathbf{v}_{m,1}^H \mathbf{H}_{m,1} \mathbf{p}_m|^2}. \tag{3.17}$$

This completes the proof for the two user per cluster case.

Assume that (3.15) holds for n users per cluster, i.e., $\Delta P_m^{(n)} = P_{\max} \sum_{l=1}^{n} \Delta\hat{\Omega}_{m,l} = (2^{\Delta R} - 1) \frac{P_{\max} 2^{\sum_{l=1}^{n} \hat{R}_{m,l}}}{\rho |\mathbf{v}_{m,1}^H \mathbf{H}_{m,1} \mathbf{p}_m|^2}$. On this basis, we consider the case with $n+1$ users. According to (3.7), the power coefficient for user $(m, n+1)$ before increasing the sum rate is given by

$$\hat{\Omega}_{m,n+1} = \frac{(2^{\hat{R}_{m,n+1}} - 1)(1 + \rho |\mathbf{v}_{m,n+1}^H \mathbf{H}_{m,n+1} \mathbf{p}_m|^2 \sum_{l=1}^{n} \hat{\Omega}_{m,l})}{\rho |\mathbf{v}_{m,n+1}^H \mathbf{H}_{m,n+1} \mathbf{p}_m|^2}. \tag{3.18}$$

After the ΔR sum rate increment, the rate of user $(m, n+1)$ remains unchanged according to Lemma 3.1. Thus, the power coefficient increment for user $(m, n+1)$ is $\Delta\hat{\Omega}_{m,n+1} = (2^{\hat{R}_{m,n+1}} - 1) \sum_{k=1}^{n} \Delta\hat{\Omega}_{m,k}$.

Accordingly, the total required extra power for the $n+1$ users can be expressed as

$$\begin{aligned}
\Delta P_m^{(n+1)} &= \Delta P_m^{(n)} + P_{\max} \Delta\hat{\Omega}_{m,n+1} \\
&= P_{\max} \sum_{k=1}^{n} \Delta\hat{\Omega}_{m,k} + P_{\max} (2^{\hat{R}_{m,n+1}} - 1) \sum_{k=1}^{n} \Delta\hat{\Omega}_{m,k} \\
&= P_{\max} 2^{\hat{R}_{m,n+1}} \sum_{k=1}^{n} \Delta\hat{\Omega}_{m,k} \\
&= 2^{\hat{R}_{m,n+1}} (2^{\Delta R} - 1) \frac{P_{\max} 2^{\sum_{l=1}^{n} \hat{R}_{m,l}}}{\rho |\mathbf{v}_{m,1}^H \mathbf{H}_{m,1} \mathbf{p}_m|^2} \\
&= (2^{\Delta R} - 1) \frac{P_{\max} 2^{\sum_{l=1}^{n+1} \hat{R}_{m,l}}}{\rho |\mathbf{v}_{m,1}^H \mathbf{H}_{m,1} \mathbf{p}_m|^2},
\end{aligned} \tag{3.19}$$

which completes the proof. ■

We observe that only the channel gains of the first user and the minimum rate requirement of all users affect the power increment for each cluster. Moreover, for smaller $\frac{P_{\max} 2^{\sum_{l=1}^{L} \hat{R}_{m,l}}}{\rho |\mathbf{v}_{m,1}^H \mathbf{H}_{m,1} \mathbf{p}_m|^2}$, less additional power is needed for increasing the sum rate by the same unit. This observation can be used for designing an iterative PA algorithm. Specifically, during each iteration, the cluster with the smallest $\frac{P_{\max} 2^{\sum_{l=1}^{L} \hat{R}_{m,l}}}{\rho |\mathbf{v}_{m,1}^H \mathbf{H}_{m,1} \mathbf{p}_m|^2}$ is selected for receiving

the additional power. On the other hand, after this cluster receives a certain amount of additional power, its sum rate $\sum_{l=1}^{L}\hat{R}_{m,l}$ increases, and it may no longer be the one with the smallest $\frac{P_{\max}2^{\sum_{l=1}^{L}\hat{R}_{m,l}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}$. This process repeats until P_{f} is fully used. This iterative algorithm is similar to the classical water-filling technique, and a closed-form solution can be obtained accordingly.

More precisely, after the initial feasible PA, we obtain $R_{m,l}=R_{m,l}^{\min}, m\in\{1,\cdots,M\}, l\in\{1,\cdots,L\}$. We consider $\frac{P_{\max}2^{\sum_{l=1}^{L}R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}$ as the initial "water" level. Furthermore, we introduce an axillary variable λ as the final "water" level. If $\lambda\leq\frac{P_{\max}2^{\sum_{l=1}^{L}R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}$, the mth cluster receives no power and remains unchanged. Otherwise, the mth cluster receives some extra power to reach the final "water" level, i.e., $\lambda=\frac{P_{\max}2^{\sum_{l=1}^{L}R_{m,l}^{\min}+\Delta R_m}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}=2^{\Delta R_m}\times\frac{P_{\max}2^{\sum_{l=1}^{L}R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}$, where ΔR_m is the rate increment. In this case, according to (3.15), the required extra power can be expressed as

$$\begin{aligned}\Delta P_m&=(2^{\Delta R_m}-1)\frac{P_{\max}2^{\sum_{l=1}^{L}R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}\\&=\lambda-\frac{P_{\max}2^{\sum_{l=1}^{L}R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}.\end{aligned}\tag{3.20}$$

Considering both cases, the required power for the mth cluster can be further expressed as

$$\Delta P_m=\left[\lambda-\frac{P_{\max}2^{\sum_{l=1}^{L}R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}\right]^{+},\tag{3.21}$$

where $x^{+}=\max(x;0)$. This provides a closed-form solution for the SE maximization, once λ is known. To attain the value of λ, we refer to the total power constraint, which should satisfy

$$\sum_{m=1}^{M}\left[\lambda-\frac{P_{\max}2^{\sum_{l=1}^{L}R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_m|^2}\right]^{+}=P_{\mathrm{rem}}.\tag{3.22}$$

The left side of the above equation is piecewise and monotonically increasing over λ. Thus, a unique value of λ exists and can be obtained by solving (3.22). Note that there

is a point-to-point mapping between λ and P_{f}, and further, λ increases with P_{f}.

Moreover, the sum rate increment for the mth cluster can be expressed as

$$\Delta R_m = \left[\log_2(\lambda) - \log_2 \left(\frac{P_{\max} 2^{\sum_{l=1}^{L} R_{m,l}^{\min}}}{\rho |\mathbf{v}_{m,1}^H \mathbf{H}_{m,1} \mathbf{p}_m|^2} \right) \right]^+. \tag{3.23}$$

The following lemma shows the optimality of the proposed SE maximization PA strategy.

Lemma 3.3. *The proposed SE maximization PA strategy maximizes the sum rate of the system.*

Proof: Assume that we have obtained the solution via the proposed PA algorithm, i.e., λ is known and so are other values, e.g., the extra power for each cluster. Now, we shift Δp power between two clusters whose final "water" level is λ. Denote the two clusters as the qth and nth cluster, respectively. For the proposed PA strategy, the sum rate increment for the qth cluster after the initial PA can be expressed as

$$\begin{aligned} \Delta R_q &= \log_2(\lambda) - \log_2 \left(\frac{P_{\max} 2^{\sum_{l=1}^{L} R_{q,l}^{\min}}}{\rho |\mathbf{v}_{q,1}^H \mathbf{H}_{q,1} \mathbf{p}_q|^2} \right) \\ &= \log_2 \left(\Delta P_q + \frac{P_{\max} 2^{\sum_{l=1}^{L} R_{q,l}^{\min}}}{\rho |\mathbf{v}_{q,1}^H \mathbf{H}_{q,1} \mathbf{p}_q|^2} \right) - \log_2 \left(\frac{P_{\max} 2^{\sum_{l=1}^{L} R_{q,l}^{\min}}}{\rho |\mathbf{v}_{q,1}^H \mathbf{H}_{q,1} \mathbf{p}_q|^2} \right). \end{aligned} \tag{3.24}$$

For ΔR_n, a similar expression can be written.

After shifting some power between two clusters, we have

$$\begin{aligned} \Delta R_q' &= \log_2(\Delta P_q + \Delta p + \frac{P_{\max} 2^{\sum_{l=1}^{L} R_{q,l}^{\min}}}{\rho |\mathbf{v}_{q,1}^H \mathbf{H}_{q,1} \mathbf{p}_q|^2}) - \log_2 \left(\frac{P_{\max} 2^{R_{q,2}^{\min}}}{\rho |\mathbf{v}_{q,1}^H \mathbf{H}_{q,1} \mathbf{p}_q|^2} \right) \\ &= \log_2(\lambda + \Delta p) - \log_2 \left(\frac{P_{\max} 2^{\sum_{l=1}^{L} R_{q,l}^{\min}}}{\rho |\mathbf{v}_{q,1}^H \mathbf{H}_{q,1} \mathbf{p}_q|^2} \right). \end{aligned} \tag{3.25}$$

Likewise, a similar equation can be written for $\Delta R_n'$.

Accordingly, we calculate the sum rate difference as follows:

$$\begin{aligned}\Delta R_{\text{sum}} &= \Delta R_q + \Delta R_n - \Delta R_q' - \Delta R_n' \\ &= \log_2(\lambda) + \log_2(\lambda) - \log_2(\lambda + \Delta p) - \log_2(\lambda - \Delta p) \\ &= \log_2(\lambda^2) - \log_2[\lambda^2 - (\Delta p)^2] > 0.\end{aligned} \tag{3.26}$$

The above equation clearly shows that shifting power between two clusters whose final "water" level is λ leads to a lower sum rate. Following the same procedure, we can also prove that this holds when shifting power from the cluster whose final "water" level equals to λ to another cluster whose final "water" level exceeds λ. This validates the optimality of the proposed scheme. ■

Now we have solved the SE maximization problem (3.12) under P_{f}. On this basis, we need to select the appropriate P_{f} to maximize the EE of the system. Consider P_{f} as the variable here, but replace it with λ owing to the point-to-point mapping between them.

Accordingly, the consumed transmit power can be rewritten as

$$\begin{aligned}P_{\text{t}} &= P_{\text{f}} \\ &= P_{\text{req}} + \sum_{m=1}^{M} \Delta P_m \\ &= P_{\text{req}} + \sum_{m=1}^{M} \left[\lambda - \frac{P_{\max} 2^{\sum_{l=1}^{L} R_{m,l}^{\min}}}{\rho |\mathbf{v}_{m,1}^H \mathbf{H}_{m,1} \mathbf{p}_m|^2} \right]^+ .\end{aligned} \tag{3.27}$$

Similarly, the sum rate can be rewritten as

$$\begin{aligned}R^{\text{sum}} &= \sum_{m=1}^{M} \sum_{l=1}^{L} R_{m,l}^{\min} + \sum_{m=1}^{M} \Delta R_m \\ &= \sum_{m=1}^{M} \sum_{l=1}^{L} R_{m,l}^{\min} + \sum_{m=1}^{M} \left[\log_2(\lambda) - \log_2 \left(\frac{P_{\max} 2^{\sum_{l=1}^{L} R_{m,l}^{\min}}}{\rho |\mathbf{v}_{m,1}^H \mathbf{H}_{m,1} \mathbf{p}_m|^2} \right) \right]^+ .\end{aligned} \tag{3.28}$$

As a result, the expression of η_{EE} can be written as follows:

$$\eta_{\mathrm{EE}} = \frac{\sum_{m=1}^{M}\sum_{l=1}^{L} R_{m,l}^{\min} + \sum_{m=1}^{M}\left[\log_2(\lambda) - \log_2\left(\frac{P_{\max}2^{\sum_{l=1}^{L} R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\mathbf{p}_m|^2}\right)\right]^+}{P_{\mathrm{c}} + P_{\mathrm{req}} + \sum_{m=1}^{M}\left[\lambda - \frac{P_{\max}2^{\sum_{l=1}^{L} R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\mathbf{p}_m|^2}\right]^+}. \tag{3.29}$$

Clearly, in (3.29), the only variable is λ, as other parameters are known. Moreover, (3.29) is a piecewise function, and its specific form depends on the interval λ lies in. To find the intervals, we arrange $\frac{P_{\max}2^{\sum_{l=1}^{L} R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^H \mathbf{H}_{m,1}\mathbf{p}_m|^2}$ in an ascending order and use h_t to denote the tth value after ordering for simplicity of notation. Since the total transmit power $P_{\max}$ is known, we can calculate the maximum index of the interval that λ can lie in according to (3.22), by setting $P_{\mathrm{rem}} = P_{\max} - P_{\mathrm{req}}$. Denote the maximum index as T, then λ can only lie in $[h_t, h_{t+1}], t = 1, \cdots, T$. Moreover, we have the following theorem:

Theorem 3.1. *For each interval $[h_t, h_{t+1}], t = 1, \cdots, T$, η_{EE} is a strictly pseudo-concave function with respect to (w.r.t.) λ.*

Proof: Once t is known, η_{EE} can be turned into

$$\eta_{\mathrm{EE}} = \frac{\sum_{m=1}^{M}\sum_{l=1}^{L} R_{m,l}^{\min} + t\log_2(\lambda) - \sum_{k=1}^{t}\log_2(h_k)}{P_{\mathrm{c}} + P_{\mathrm{req}} + t\lambda - \sum_{k=1}^{t} h_k}. \tag{3.30}$$

It can be seen that the numerator is a strictly concave function over λ, while the denominator is an affine mapping over λ. Thus, η_{EE} is a strictly pseudo-concave function w.r.t. λ [19, Proposition 6]. ■

For $\lambda \in [h_t, h_{t+1}]$, as η_{EE} is a strictly pseudo-concave function w.r.t. λ, η_{EE} admits a unique maximizer, which is obtained either at the unique root of the equation $\frac{\partial \eta_{\mathrm{EE}}}{\partial \lambda} = 0$ or at the two boundary points h_t or h_{t+1} [19, Proposition 5]. Denote this maximizer as η_{EE}^t. Likewise, when λ lies in $[h_k, h_{k+1}], k \neq t$, denote the unique maximizer as η_{EE}^k. As η_{EE} belongs to two different functions for these two intervals, we cannot determine the

comparative values of these two maximizers analytically. Instead, an explicit comparison has to be done, i.e., $\max\left\{\eta_{EE}^{t}, \eta_{EE}^{k}\right\}$. As the total number of intervals is T, we need to obtain the maximizer in each interval and select the maximum for η_{EE}, which can be expressed as

$$\eta_{\mathrm{EE}}^{\max} = \max\left\{\eta_{\mathrm{EE}}^{1}, \cdots, \eta_{\mathrm{EE}}^{T}\right\}. \tag{3.31}$$

So far, we have derived the solution for maximizing the EE of the system. We summarize the procedures in Algorithm 1. Moreover, the following theorem proves its optimality.

Theorem 3.2. *The derived solution achieves the maximum EE for the system.*

Proof: According to Lemma 3.3, for any given total power, the proposed solution maximizes the SE of the system by appropriately allocating power across clusters and inside each cluster. Then, Theorem 3.1 guarantees that the EE is maximized for each feasible interval. As (3.31) selects the maximum value from all these maximizers, this selected maximum value is the global optimum. ■

Algorithm 1 Proposed EE Maximization PA Algorithm

Initialize parameters.
  $P_{\max}$, $R_{m,l}^{\min}$, $\rho|\mathbf{v}_{m,l}^{H}\mathbf{H}_{m,l}\mathbf{p}_{m}|^{2}, l \in \{1, \cdots, L\}$
Calculate:
  $\mathbf{H} \leftarrow sort\left(\frac{P_{\max}2^{\sum_{l=1}^{L} R_{m,l}^{\min}}}{\rho|\mathbf{v}_{m,1}^{H}\mathbf{H}_{m,1}\mathbf{p}_{m}|^{2}}\right)$;
  $h_t \leftarrow \mathbf{H}(t)$;
  $\lambda^{\max} \leftarrow \sum_{t=1}^{M}[\lambda - h_t]^{+} = P_{\max} - P_{\mathrm{req}}$;
  $T \leftarrow \lambda^{\max} \in [h_T, h_{T+1}]$;
  $\eta_{\mathrm{EE}}^{t} \leftarrow \max\left\{\eta_{\mathrm{EE}}(\frac{\partial \eta_{\mathrm{EE}}}{\partial \lambda} = 0), \eta_{\mathrm{EE}}(h_t), \eta_{\mathrm{EE}}(h_{t+1})\right\}, t \in \{1, \cdots, T-1\}$;
  $\eta_{\mathrm{EE}}^{T} \leftarrow \max\left\{\eta_{\mathrm{EE}}(\frac{\partial \eta_{\mathrm{EE}}}{\partial \lambda} = 0), \eta_{\mathrm{EE}}(h_T), \eta_{\mathrm{EE}}(\lambda^{\max})\right\}$;
  $\eta_{\mathrm{EE}}^{\max} \leftarrow \max\left\{\eta_{\mathrm{EE}}^{1}, \cdots, \eta_{\mathrm{EE}}^{T}\right\}$;
end

3.4.2 User Admission when Problem (3.9) is Infeasible

When (3.9) is infeasible, admitting as many users as possible is a more reasonable goal, when compared with EE maximization. The user admission problem can be formulated as

$$\max_{\Omega_{m,l}} \quad \sum_{m=1}^{M}\sum_{l=1}^{L} x_{m,l} \tag{3.32a}$$

$$\text{s.t.} \quad R_{m,l} \geq R_{m,l}^{\min} x_{m,l}, \tag{3.32b}$$

$$\sum_{m=1}^{M}\sum_{l=1}^{L} \Omega_{m,l} \leq 1, \tag{3.32c}$$

$$x_{m,l} \in \{0,1\}, \tag{3.32d}$$

where $x_{m,l}$ is the binary decision variable indicating whether user (m,l) is admitted or not.

In [12], under the assumption of equal power for each cluster, we propose a greedy user admission algorithm, which admits users one by one following the descending order of their channel gains within each cluster. In this chapter, as power can be transferred among clusters, user admission should be conducted globally. Based on the observation that the users with better channel gains own higher priority than their counterparts in each cluster due to SIC, we still admit users within each cluster following the descending order of their channel gains. Furthermore, with multi-clusters in the system, we also need to determine the order for admitting users across clusters. This can be done by comparing the required power for satisfying the QoS of each user in different clusters, and select the one with the minimum power consumption during each user admission process.

More exactly, the user admission is conducted iteratively as follows: during each iteration, we first select the user with the best channel gain from each cluster; among these selected users, the required power is calculated with considering the interference from the already admitted users; then, the user with the minimum required power is chosen to

be admitted; if the total remaining power exceeds the required power for admitting this user, the selected user is admitted and eliminated from the candidates; besides, the total remaining power is updated; otherwise, the process terminates; the process repeats until no further user can be admitted.

Theorem 3.3. *The proposed scheme maximizes the number of admitted users when the users' QoS requirements in each cluster satisfy the following conditions:*

$$R_{m,k}^{\min} \leq R_{m,n}^{\min}, \forall k \in \{1, \ldots, l\}, n \in \{l+1, \ldots, L\}, \tag{3.33}$$

where l represents the total number of admitted users in the mth cluster under the proposed scheme.

Proof: Refer to Appendix. ■

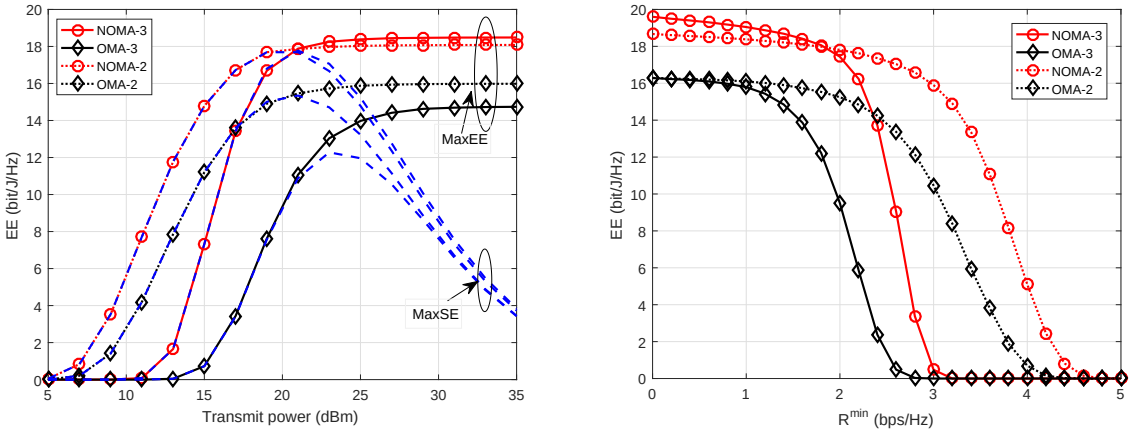
Corollary 3. *The proposed user admission scheme maximizes the number of admitted users when the SINR thresholds of the users in each cluster satisfy the following conditions:*

$$R_{m,1}^{\min} \leq \cdots \leq R_{m,L}^{\min}. \tag{3.34}$$

Particularly, when the QoS requirements of the users are equal, the proposed user admission scheme is optimal in terms of both sum rate and number of admitted users.

Proof: When (3.34) is satisfied, it can be easily proved that (3.33) holds for any l. Thus, the proposed scheme maximizes the number of admitted users. If the QoS requirements of the users are equal, it can be easily inferred that maximizing the number of admitted users also leads to the maximization of the sum rate. ■

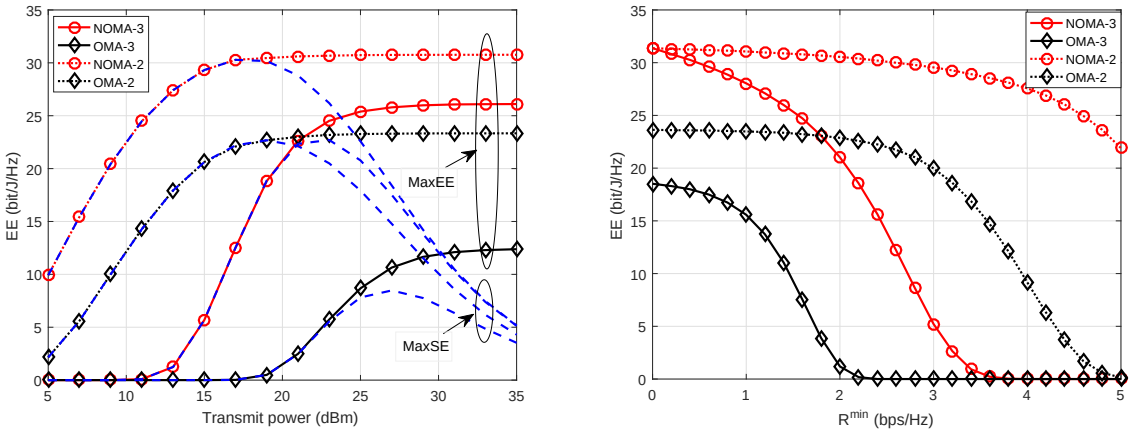
Lemma 3.4. *The complexity of the proposed user admission algorithm is $O(M^2L)$.*



(a) How EE varies with the transmit power: $R^{\min} = 2$ bps/Hz

(b) How EE varies with the minimum rate requirement: $P_{\mathrm{t}} = 20$ dBm

Fig. 3.1: Scenario 1: $d_1 = d_2 = d_3 = 80$ m.



(a) How EE varies with the transmit power: $R^{\min} = 2$ bps/Hz

(b) How EE varies with the minimum rate requirement: $P_{\mathrm{t}} = 20$ dBm

Fig. 3.2: Scenario 2: $d_1 = 40\ m, d_2 = 80\ m, d_3 = 120\ m$.

Proof: The proposed user admission algorithm admits users one by one following the ascending order of the required power for satisfying their QoS requirements, which requires $O(ML)$ operations. During each user admission, the main complexity comes from the operation of selecting the minimum value across all clusters, which requires $O(M)$ operations. In all, the complexity of the proposed user admission algorithm is $O(M^2L)$. ■

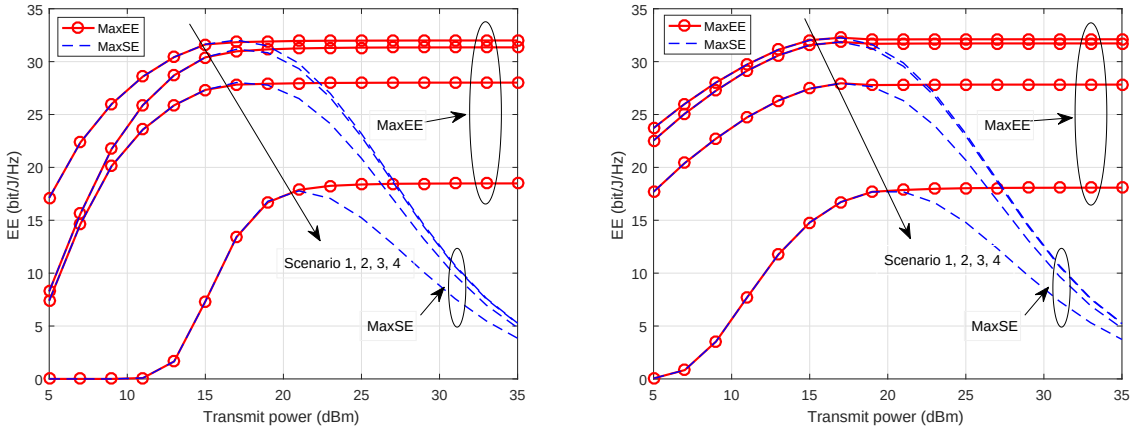
Table 3.1: Simulation Parameters.

Parameters	**Value**
Number of antennas	$M = 3, N = 3$
Channel bandwidth	10 [MHz]
Thermal noise density	-174 [dBm/Hz]
Path-loss model	$120 + 30 \log_{10}(d)$, d in kilometer

3.5 Simulation Results

In this section, simulations are conducted to verify the performance of the proposed PA strategy and user admission scheme. The specific values of the adopted simulation parameters are summarized in Table 3.1 [12]. All results are obtained by averaging over 10^4 random trials, unless mentioned otherwise. Particularly, in the case when the total transmit power cannot support the QoS requirements for all users, the EE of these trials is set to zero since the objective is not EE maximization.

First, the effectiveness of the proposed energy-efficient PA strategy is evaluated. To compare NOMA with conventional OMA, OMA with equal degrees of freedom for each user is adopted as the baseline algorithm. Note that OMA can be considered as a special case of NOMA, with one user in each cluster. The energy-efficient PA for OMA can be attained by employing the proposed energy-efficient PA strategy for NOMA with some minor adjustment, e.g., now the cluster number becomes $M \times L$. The above energy-efficient PA strategies are denoted as "MaxEE". As a baseline algorithm, the PA strategy



(a) Three users per cluster

(b) Two users per cluster

Fig. 3.3: EE versus total power available at the BS, for different cases of user locations.

that consumes full power to maximize the SE of the system is also presented, which is denoted as "MaxSE". This "MaxSE" PA for NOMA can be obtained by employing the proposed water-filling sum rate maximization algorithm. In terms of the "MaxSE" PA for OMA, it can be achieved by employing the classical water-filling algorithm.

To show how EE varies as the number of users in each cluster increases, two scenarios with different distances are presented in Figs. 3.1 and 3.2, in which "-3" and "-2" mean three and two users per cluster, respectively. For each scenario, how EE varies with the total transmit power and minimum rate requirement is presented. According to Figs. 3.1 and 3.2, NOMA achieves higher EE than OMA for both two and three user cases, respectively.

Specifically, subfigures 3.1(a) and 3.2(a) show how EE varies with the transmit power, in which the dashed lines in both figures denote the "MaxSE", while all other lines represent the "MaxEE". Clearly, under low transmit power, "MaxSE" equals "MaxEE", and both grow with the transmit power. As the transmit power reaches a certain threshold, further increase in the transmit power does not yield a higher EE, and thus, "MaxEE" remains stable, while "MaxSE" decreases. This indicates the necessity of employing energy-

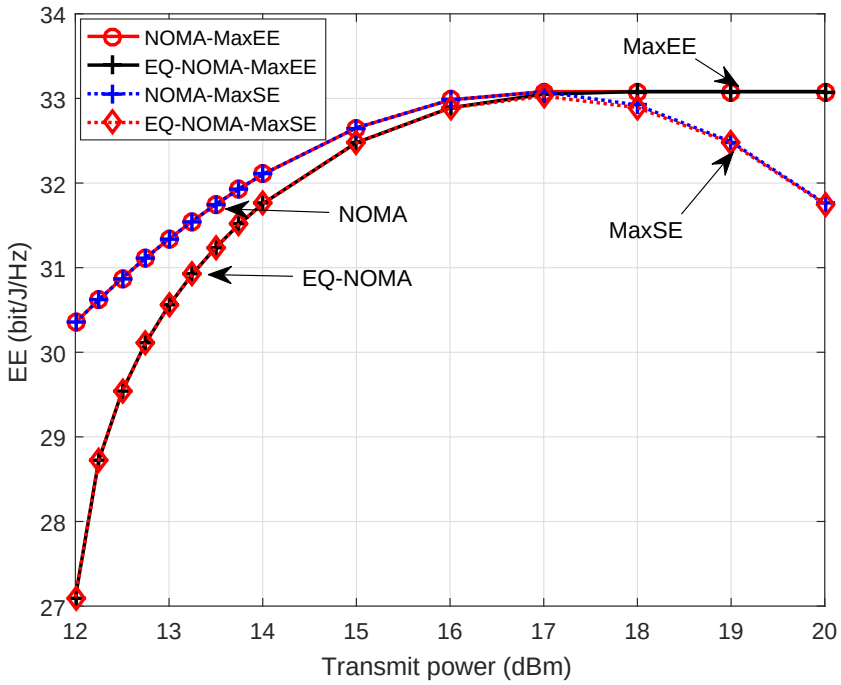


Fig. 3.4: EE versus total power available at the BS, for NOMA and EQ-NOMA.

efficient PA, especially under high transmit power. In scenario 1, under low transmit power, NOMA-2 achieves higher EE compared with NOMA-3. However, under high transmit power, an opposite result can be observed. This can be explained by the fact that under low transmit power, it is more difficult to satisfy the QoS for three users. On the other hand, under high power, more users lead to a higher diversity, which increases the EE. In contrast, in scenario 2, NOMA-2 always attains higher EE than its counterpart. This is due to the fact as the distance difference between the users increases, it costs more energy to admit an extra user. Thus, even under high transmit power, the benefit introduced by the diversity is not enough to compensate the energy required for admitting the extra user. Combining the two scenarios, we can conclude that whether admitting more users yields a higher EE depends on the transmit power level and the distance difference between the users.

Subfigures 3.1(b) and 3.2(b) show how EE varies with $R^{\min}$. It can be seen that

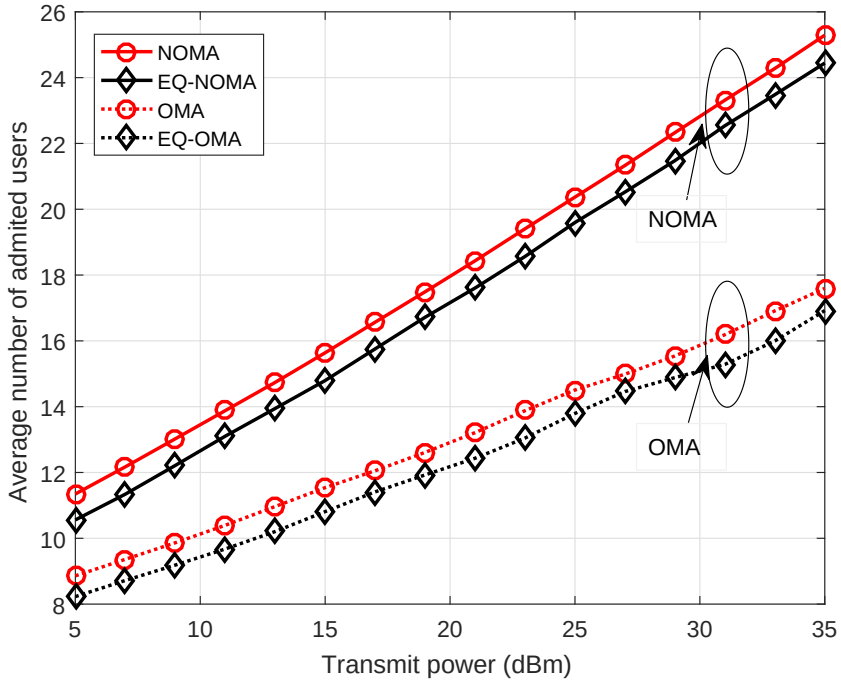


Fig. 3.5: Average number of admitted users versus transmit power: number of requesting users per cluster is 15; $R^{\min} = 2$ bps/Hz.

EE decreases with $R^{\min}$. More exactly, in scenario 1, NOMA-3 achieves higher EE than NOMA-2 under low $R^{\min}$, and vice versa. This can be explained by connecting $R^{\min}$ with the transmit power, i.e., lower $R^{\min}$ has the same impact on EE as higher transmit power. In contrast, in scenario 2, NOMA-2 always achieves higher EE than NOMA-3, which agrees with subfigure (a).

Results in Figs. 3.1 and 3.2 indicate that the distance has an impact on EE; accordingly, in Fig. 3.3, further analysis on this is provided under four scenarios. Scenario 1: $d_1 = 60$ m, $d_2 = 50$ m, $d_3 = 40$ m, $(d_1 + d_2 + d_3)/3 = 50$ m. Scenario 2: $d_1 = 70$ m, $d_2 = 55$ m, $d_3 = 40$ m, $(d_1 + d_2 + d_3)/3 = 55$ m. Scenario 3: $d_1 = 60$ m, $d_2 = 55$ m, $d_3 = 50$ m, $(d_1 + d_2 + d_3)/3 = 55$ m. Scenario 4: $d_1 = 80$ m, $d_2 = 80$ m, $d_3 = 80$ m, $(d_1 + d_2 + d_3)/3 = 80$ m. Obviously, the larger the distance, the lower the achieved EE. Furthermore, comparing scenarios 2 and 3, we can conclude that the channel gain of the

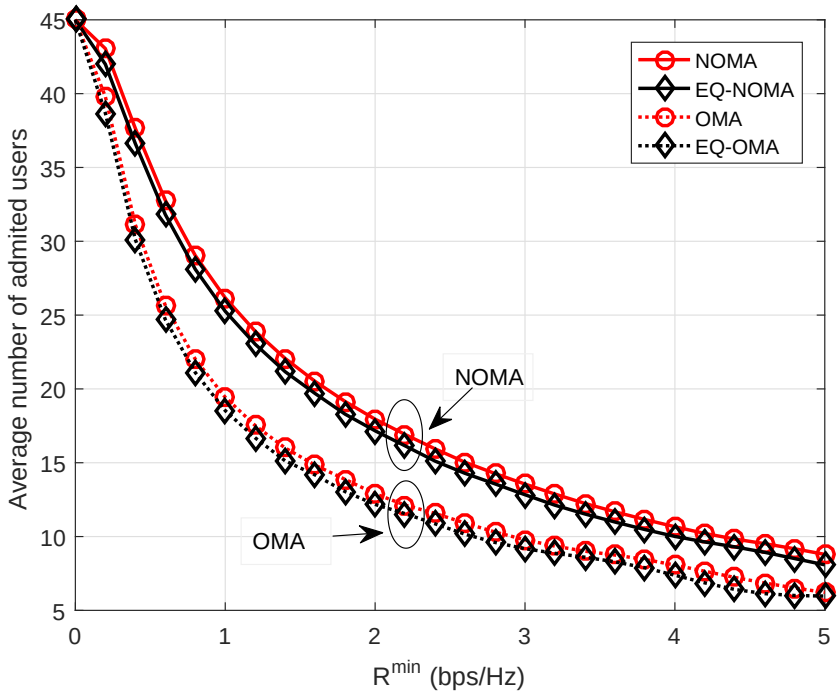


Fig. 3.6: Average number of admitted users versus $R^{\min}$: number of requesting users per cluster is 15; $P_{\mathrm{t}} = 20$ dBm.

strongest user plays a vital role in EE, which fits our observation in Lemma 3.2. On the other hand, by comparing three and two user cases for scenario 2, it implies that the distance difference between users has a larger impact on the multi-user case, especially under lower transmit power. To conclude, not only the average distance, but also the distance of the strongest user plays an important role in EE. Moreover, the distance difference affects EE more for the three user case under low transmit power.

Figure 3.4 compares EE achieved by the proposed PA strategy with that achieved by the algorithm in [17], in which equal power is assigned to each cluster, and thus is denoted as "EQ-NOMA". Further, for both algorithms, both "MaxEE" and "MaxSE" are plotted. It can be seen that under low transmit power, the proposed PA strategy achieves higher EE than the one in [17], which validates the necessity of applying global PA. On the other hand, under high transmit power, their performance is the same. This can be

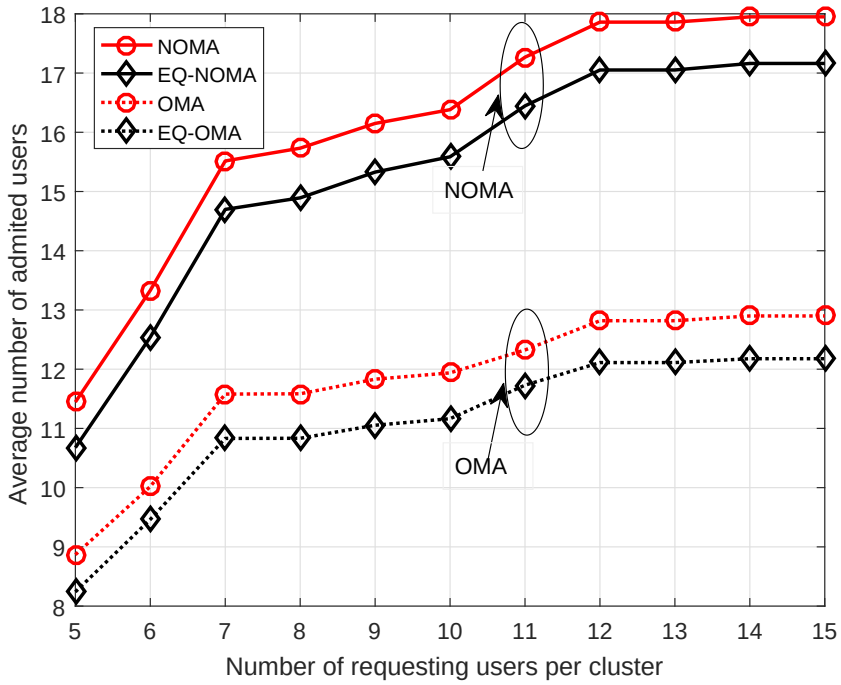


Fig. 3.7: Average number of admitted users versus number of requesting users per cluster: $R^{\min} = 2$ bps/Hz; $P_{\mathrm{t}} = 20$ dBm.

explained by the fact that under high transmit power, the equally divided power is enough for EE maximization, and thus, allowing power to be transferred among clusters brings no benefit.

Figs. 3.5-3.7 show the performance of the proposed user admission scheme, which is denoted as "NOMA". As a baseline algorithm, we consider the NOMA scheme in [12], which assigns equal power to each cluster, and is denoted as "EQ-NOMA". To compare NOMA with conventional OMA, OMA with PA across clusters and OMA with equal power per cluster are presented, denoted as "OMA" and "EQ-OMA", respectively. According to Figs. 3.5-3.7, it can be seen that NOMA outperforms OMA in terms of the number of admitted users versus transmit power, minimum rate requirement, and number of requesting users. Moreover, for both NOMA and OMA, allowing power to be transferred among clusters leads to a larger number of admitted users. In addition, it is clear

that the average number of admitted users grows with the transmit power, but decreases with $R^{\min}$. Furthermore, it also increases with the number of requesting users per cluster. This is due to the fact that when more users are requesting the service, it is more likely that more users will have a better channel gains, yielding a lower power to satisfy their minimum rate requirements. As the total power is fixed, more users can be admitted accordingly.

3.6 Conclusion

In this chapter, the EE maximization problem has been studied for a multi-cluster multi-user MIMO-NOMA system under a QoS constraint for each user. An optimal PA strategy has been proposed to solve the considered EE maximization problem when it is feasible. A low complexity user admission protocol has been proposed otherwise, which admits users one by one following the ascending order of the required power for satisfying the QoS requirements. Numerical results have shown that the proposed PA strategies outperform OMA and equal power NOMA in terms of both EE and the number of admitted users, which verifies their effectiveness. In addition, the EE of the NOMA system mainly depends on the channel condition of the first user, and it is necessary to apply an energy-efficient PA strategy, especially at high transmit power. On the other hand, whether more users leads to increased EE depends on the transmit power level and users' channel gain difference.

Appendix

Proof of Theorem 3

Proof: The user admission in each cluster is first considered. In the following, I will prove through contradiction that the proposed scheme maximizes the number of admitted users in each cluster.

Consider the case in which only l users can be admitted to the mth cluster when employing the proposed user admission scheme. Suppose there exists an alternate scheme, which also admits l users, but replaces the kth user with the nth user as one admitted user, $k \in \{1, \ldots, l\}, n \in \{l+1, \ldots, L\}$. In this case, it seems that the alternate scheme transfers the power of the kth user to the nth user. Moreover, from the $(k+1)$th user to the lth user, the required power for satisfying their QoS requirements decreases, as the interference from the kth user is removed. This reduced power can also be considered to be transferred to the nth user. According to the remark from Lemma 3.1, a lower sum rate is achieved by transferring power from the strong users to the weak users. Since all other users' rates remain the same, the achievable rate of the nth user must be lower than that of the kth user, $R_{m,n} \leq R_{m,k}^{\min}$. On the other hand, $R_{m,k}^{\min} \leq R_{m,n}^{\min}$. Therefore, $R_{m,n} \leq R_{m,n}^{\min}$, which indicates that more power is needed to satisfy the QoS requirement of the nth user. This shows that the proposed scheme requires the minimum power when there is one replacement between the users. Following the same procedure, the conclusion can be easily extended to the case in which there exist multiple replacements of the users, which means that the proposed scheme requires the minimum power for admitting l users, i.e., $\Omega_{\text{sum}} \leq \Omega_{\text{sum}}^{\text{alt}}$, where Ω_{sum} and $\Omega_{\text{sum}}^{\text{alt}}$ are the total power coefficients of admitting l users for the proposed scheme and the alternate one, respectively.

Suppose the alternate scheme can admit an extra user, denoted as a_{l+1}. Without loss of generality, the channel gain of this user is assumed to be the lowest. Note that this

assumption does not add an extra constraint since we can simply exchange it with the one of the lowest channel gain, and consider the latter as the extra admitted user. According to (3.7), $\Omega^{\text{alt}}_{m,a_{l+1}} \geq (2^{R^{\min}_{m,a_{l+1}}} - 1)\left(\Omega^{\text{alt}}_{\text{sum}} + \frac{1}{\rho|\mathbf{v}^H_{m,a_{l+1}}\mathbf{H}_{m,a_{l+1}}\mathbf{p}_m|^2}\right)$. In addition, admitting the a_{l+1} to the proposed scheme requires $\Omega_{m,a_{l+1}} = (2^{R^{\min}_{m,a_{l+1}}} - 1)\left(\Omega_{\text{sum}} + \frac{1}{\rho|\mathbf{v}^H_{m,a_{l+1}}\mathbf{H}_{m,a_{l+1}}\mathbf{p}_m|^2}\right)$. As $\Omega_{\text{sum}} \leq \Omega^{\text{alt}}_{\text{sum}}$, it can be obtained that $\Omega_{m,a_{l+1}} + \Omega_{\text{sum}} \leq \Omega^{\text{alt}}_{m,a_{l+1}} + \Omega^{\text{alt}}_{\text{sum}}$. Thus, this extra user can also be admitted to the proposed scheme, which conflicts with the proposition that only l users can be admitted by the proposed scheme.

With multi-clusters, since the proposed scheme selects the user with the minimum required power across clusters during each iteration, this clearly yields the maximum number of admitted users. ■

References

[1] S. M. R. Islam, M. Zeng, and O. A. Dobre, "NOMA in 5G systems: Exciting possibilities for enhancing spectral efficiency," *IEEE 5G Tech. Focus,* vol. 1, no. 2, May 2017. [Online]. Available: http://5g.ieee.org/tech-focus.

[2] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surv. Tuts.*, vol. pp, no. 99, pp. 1–1, Oct. 2016.

[3] L. Dai, B. Wang, Y. Yuan, S. Han, C. L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[4] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[5] L. Song, Y. Li, Z. Ding, and H. V. Poor, "Resource management in non-orthogonal multiple access networks for 5G and beyond," *IEEE Network*, vol. 31, no. 4, pp. 8–14, Jul. 2017.

[6] 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 Veh. Technol. Conf.*, Dresden, Germany, Jun. 2013, pp. 1–5.

[7] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, UK: Cambridge University Press, 2005.

[8] 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, Dec. 2014.

[9] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, Jun. 2016.

[10] 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, Aug. 2017.

[11] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[12] 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. Select. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.

[13] B. Wang, L. Dai, Z. Wang, N. Ge, and S. Zhou, "Spectrum and energy efficient beamspace MIMO-NOMA for millimeter-wave communications using lens antenna array," *IEEE J. Select. Areas Commun.*, to appear, 2017.

[14] V. Nguyen, H. Tuan, T. Duong, H.V., Poor, and O. Shin, "Precoder design for signal superposition in MIMO-NOMA multicell networks," *IEEE J. Select. Areas Commun.*, vol. PP, no. 99, pp. 1–1, Jul. 2017.

[15] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722–3732, Sep. 2016.

[16] Y. Zhang, H. M. Wang, T. X. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2852–2857, Mar. 2017.

[17] W. M. Hao, et al., "Energy-efficient power allocation in millimeter wave massive mimo with non-orthogonal multiple access," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 782–785, Dec 2017.

[18] M. Zeng, G. I. Tsiropoulos, O. A. Dobre, and M. H. Ahmed, "Power allocation for cognitive radio networks employing non-orthogonal multiple access," in *Proc. IEEE Global Commun. Conf.*, Washington DC, USA, Dec. 2016.

[19] A. Zappone, P. Lin, and E. Jorswieck, "Energy efficiency in secure multi-antenna systems," *IEEE Trans. Signal Process.,* submitted for publication. [Online]. Available: http://arxiv.org/abs/1505.02385.

Chapter 4

Energy-Efficient Joint User-RB Association and Power Allocation for Uplink Hybrid NOMA-OMA

4.1 Abstract

In this chapter, energy efficient resource allocation is considered for an uplink hybrid system, where NOMA is integrated into OMA. To ensure the quality of service for the users, a minimum rate requirement is pre-defined for each user. An EE maximization problem is formulated by jointly optimizing the user clustering, channel assignment and power allocation. To address this hard problem, a many-to-one bipartite graph is first constructed considering the users and resource blocks (RBs) as the two sets of nodes. Based on swap matching, a joint user-RB association and power allocation scheme is proposed, which converges within a limited number of iterations. Moreover, for the power allocation under a given user-RB association, the feasibility condition is first derived. If feasible, a low-complexity algorithm is proposed, which obtains optimal EE under any

SIC order and an arbitrary number of users. In addition, for the special case of two users per cluster, analytical solutions are provided for the two SIC orders, respectively. These solutions shed light on how the power is allocated for each user to maximize the EE. Numerical results are presented, which show that the proposed joint user-RB association and power allocation algorithm outperforms other hybrid multiple access based and OMA-based schemes.

4.2 Introduction

NOMA has been considered as a promising candidate for the fifth generation (5G) and beyond 5G cellular networks [2–7]. The key idea of NOMA is to serve multiple users simultaneously over the same radio resources. The introduced inter-user interference is mitigated by employing SIC at the receiver. Downlink NOMA has been extensively studied so far. Some works target sum rate maximization and show that higher spectral efficiency (SE) can be achieved by NOMA when compared with conventional OMA [8–13]. Other works study EE maximization and show that NOMA can also deliver higher EE than OMA [14–18]. In addition, NOMA has also been applied to downlink celluar machine-to-machine (M2M) communication, and it is shown that improved outage probability can be achieved by NOMA when compared with OMA [19].

While uplink NOMA has been less studied compared with downlink NOMA, it has been gaining more attention recently [20–27]. In [20] and [21], system-level throughput performance is studied, and it is shown that compared with OMA, enhanced performance can be obtained by NOMA through proportional fair-based scheduling and fractional transmission power control. [22] proposes to incorporate multi-level received power and sequence grouping into existing NOMA schemes, and shows that the proposed scheme can support larger connectivity and higher reliability. In terms of connectivity, since

SIC is conducted at the BS, which is less complexity- and energy-constrained, uplink NOMA can support more users than the downlink case. This makes it a promising candidate for providing massive connectivity for the Internet-of-Things (IoT) [23,24]. [23] proposes a non-orthogonal random access (NORA) scheme based on SIC to alleviate the access congestion problem facing IoT. Analytical and simulation results verify the superiority of NORA over OMA in terms of the preamble collision probability, access success probability, and throughput. [24] also considers a random NOMA strategy for massive IoT, and derives system stability conditions for the maximum packet arrival rate with and without QoS guarantee. However, in the above works on random access, since the focus is on system stability, collision probability, and throughput, quite simple power allocation (PA) algorithms are used, e.g., in [23], the power back-off parameter is simply based on the index of the SIC order.

Owing to the vital role of PA in uplink NOMA, such as affecting the rate distribution among users, and determining their channel access, it deserves further study. In uplink NOMA, the SIC receiver requires diverse arrived power levels to distinguish user signals. This is quite different from OMA systems, in which an equal arrived power is desired by the BS to provide uniform QoS. In [25], joint power control and beamforming is studied to maximize the system sum rate for millimeter-wave communications. A sub-optimal solution is proposed, and simulation results show that the proposed solution achieves a close-to-bound uplink sum-rate performance. However, it only applies to single carrier system with two users. The authors in [26] consider a multi-carrier system, in which each subcarrier can support multiple users. A greedy user clustering algorithm is first proposed based on users' channel gains. Then, closed-form power allocation solutions are derived. However, [26] is based on the strong and impractical assumption that each user has the same channel gain over different subcarriers. This is overcome by [27], in which the authors first derive the optimal PA under given channel assignment, and then propose

a low-complexity joint channel assignment and power allocation using maximum weighted independent set in graph theory. Nonetheless, the proposed solution in [27] only supports two users on a subcarrier.

The aforementioned PA schemes are for SE maximization. With EE becoming a major concern for 5G, studying PA under EE is of importance, especially for power-constrained user equipments [28]. The energy minimization of NOMA for uplink cellular M2M communications is studied in [29], where it is shown that transmitting with minimum rate and full time minimizes the energy consumption. In [30], energy-efficient PA for uplink mmWave massive MIMO system with NOMA is studied, and it is shown that NOMA can deliver superior EE when compared with OMA. Note that [30] also only allows two users to form a NOMA cluster. Different from previous works, in this chapter, The EE of an uplink hybrid system with NOMA integrated into OMA (HMA) is studied to support a larger number of IoT devices. The reasons for adopting the HMA system instead of simply applying NOMA among all IoT devices are as follows: 1) the IoT devices may not be able to process over the whole available bandwidth; 2) the delay introduced in decoding the superposed signals may be too large for the IoT-based application; note that for both NOMA and OMA, the total number of decoding is the same, but NOMA has to be done sequentially, while OMA can be done in parallel; 3) the error propagation in SIC can become severe as the number of users increases. The system objective is to maximize the EE of the considered system under an arbitrary number of users, each with a minimum rate requirement.

The considered EE maximization problem requires a joint consideration of user clustering, channel assignment and PA, which is non-convex and challenging to handle. To tackle it, a many-to-one bipartite graph is first constructed, considering the users and resources blocks (RBs) as the two sets of nodes. Then, based on swap matching, a joint user-RB association and PA algorithm is proposed, which is guaranteed to converge. Moreover,

regarding the power allocation under a given user-RB association, its feasibility conditions are first derived. If feasible, the considered problem is shown to be pseudo-concave and a low-complexity algorithm is proposed, which can obtain optimal EE for any SIC order and an arbitrary number of users. Moreover, to further shed light on how the power is allocated for each user to maximize the EE, analytical solutions are derived for the special case of two users per cluster for the two SIC orders, respectively, by exploiting the property of pseudo-concave function. Extensive numerical simulations are performed, which validate the superiority of the proposed joint user-RB association and PA scheme over other HMA- and OMA-based algorithms.

The rest of the chapter is organized as follows: Section 4.2 introduces the system model and problem formulation; Section 4.3 presents the proposed joint user-RB association and power allocation scheme; Section 4.4 shows the proposed low-complexity optimal PA algorithm under a given user-RB association; Section 4.5 discusses the special case of two users per cluster; Section 4.6 shows the simulation results; Section 4.7 finally draws the conclusions.

4.3 System Model and Problem Formulation

4.3.1 System Model

In this chapter, uplink is considered, in which a set of users denoted by $\mathcal{U} = \{1, \cdots, U\}$ require to simultaneously access the BS. The overall system bandwidth is B Hz, which is equally divided into M resource blocks (RBs), each with $\frac{B}{M}$ Hz. It is assumed that each RB can accommodate multiple users by employing NOMA, while each user can access only one RB [14, 26, 27]. The considered scheme has the flavor of both NOMA and OMA techniques, and is thus, referred to as HMA. Users sharing the same RB form a cluster. Considering user fairness, the number of users accommodated by the mth RB

is given by $L_m = \lceil \frac{U}{M} \rceil - 1$ or $\lceil \frac{U}{M} \rceil, \forall m \in \{1, \cdots, M\}$, and $\sum_{m=1}^{M} L_m = U$. Here, we assume that user-RB association is already performed for the sake of presentation, i.e., user $(m, l), l \in \{1, \cdots, L_m\}$ means the lth user in the mth RB. The way of conducting user-RB association will be presented in the next section. Let us denote the channel of user (m, l) as $h_{m,l}$, which is characterized by large scale path-loss and small scale Rayleigh fading. Without loss of generality, we also assume that the users' channels are arranged in a descending order on each RB: $|h_{m,1}| \geq \cdots \geq |h_{m,L_m}|, \forall m \in \{1, \cdots M\}$. According to the NOMA protocol, the received signal at the BS on RB m is given by

$$y_m = \sum_{l=1}^{L_m} \sqrt{P_{m,l}} h_{m,l} s_{m,l} + n_m, \tag{4.1}$$

where $s_{m,l}$ denotes the signal transmitted from the lth user over the mth RB, satisfying $\mathbb{E}(|s_{m,l}|^2) = 1$. In addition, $P_{m,l}$ denotes the corresponding transmit power, satisfying $P_{m,l} \leq P_{m,l}^{\max}$, where $P_{m,l}^{\max}$ is the maximum transmit power for user (m, l). n_m denotes the additive white Gaussian noise (AWGN) at the mth RB, which is of zero-mean and variance σ^2. Different from downlink, all received signals at the BS are desired signals in uplink, although there is multiuser interference.

In downlink, the SIC order is fixed and follows the ascending order of the channel gains, i.e., the users with lower channel gains are decoded first and removed. In contrast, in uplink, the SIC order can be flexible as all received signals at the BS are desired signals, e.g., the BS can choose to decode the user in an arbitrary order. However, regardless of that, in order to apply SIC and decode signals at the BS, PA should be fully exploited such that the distinctness among various signals is maintained. As a result, conventional PA strategies for OMA (typically intended to equalize the received signal powers for all users) are not suitable for uplink NOMA systems. For the sake of analysis, here we assume that the SIC order which decodes user 1 first is employed at the BS. Note that the developed analytical results can be easily extended to other SIC orders. Also, for the special case of

two users per cluster, the corresponding two SIC orders are explicitly studied later in the chapter. According to the NOMA protocol, the achievable rate (bit/s/Hz) for user (m, l) can be expressed as

$$R_{m,l} = \log_2 \left(1 + \frac{P_{m,l}|h_{m,l}|^2}{\sum_{k=l+1}^{L_m} P_{m,k}|h_{m,k}|^2 + \sigma^2} \right), \tag{4.2}$$

where $\sum_{k=l+1}^{L_m} P_{m,k}|h_{m,k}|^2$ denotes the inter-user interference after SIC. Particularly, when $k = L_m$, we assume $\sum_{k=L_m+1}^{L_m} P_{m,k}|h_{m,k}|^2 = 0$, i.e., user (m, L_m) receives no interference from other users.

4.3.2 Problem Formulation

The objective is to maximize the EE of the considered system while guaranteeing a minimum QoS for each user, i.e., $R_{m,l} \geq R_{m,l}^{\min}$. Note that in uplink, since users are constrained by their own individual maximum transmit power, and only receive interference from users in the same cluster due to orthogonal resources assigned to each cluster, each user may not concern the whole system EE, but its own cluster EE. Nonetheless, with multiple users and RBs, we need to consider the system EE by appropriately pairing the users and assigning the RBs.

The EE for each cluster is defined as the ratio of the achievable cluster sum rate over the total consumed power [15]. The achievable cluster sum rate is given by $R_m^{\text{sum}} = \sum_{l=1}^{L_m} R_{m,l}$, while the total power consumption includes two parts: the fixed circuit power consumption P_m^{f} and the flexible transmit power $P_m^{\text{t}} = \sum_{l=1}^{L_m} P_{m,l}$. Therefore, the EE for the mth cluster is given by

$$\eta_m^{\text{EE}} = \frac{R_m^{\text{sum}}}{P_m^{\text{f}} + P_m^{\text{t}}}. \tag{4.3}$$

Accordingly, the considered problem can be formulated as

$$\max_{P_{m,l}} \ \eta_{\mathrm{S}}^{\mathrm{EE}} \tag{4.4a}$$

$$\text{s.t.} \ \ R_{m,l} \geq R_{m,l}^{\min}, \forall m, l \in \{1, \cdots, L_m\} \tag{4.4b}$$

$$P_{m,l} \leq P_{m,l}^{\max}, \forall m, l \in \{1, \cdots, L_m\}, \tag{4.4c}$$

where $\eta_{\mathrm{S}}^{\mathrm{EE}} = \sum_{m=1}^{M} \eta_m^{\mathrm{EE}}$ denotes the system EE. (4.4b) and (4.4c) denote the QoS requirement and the transmit power constraint for each user, respectively.

4.4 Joint User-RB Association and Power Allocation (PA)

As the considered system is hybrid, it is required to associate the users with the RBs, and allocate the power. However, deriving an optimal joint user-RB association and PA scheme is challenging owing to the intra-cluster interference among users. Indeed, changing the association of a user from one RB to another not only influences this user, but also affects the other users in these RBs. Moreover, the objective function (4.4a) is non-convex, which makes it difficult to derive conditions for optimality.

4.4.1 Proposed Algorithm

To develop a low-complexity joint user-RB association and PA algorithm, the users and RBs are considered as two sets of nodes in a bipartite graph. Then, the objective is to match the users to the RBs and allocate power appropriately such that the EE can be maximized. First, a matching is defined as an assignment of RBs to users as follows.

Definition 1: Given two disjoint sets, $\mathcal{U} = \{1, \cdots, U\}$ of the users, and $\mathcal{M} = \{1, \cdots, M\}$ of the RBs, a many-to-one matching Φ is a mapping from the set $\mathcal{U} \cup \mathcal{M}$ into the set of

all subsets of $\mathcal{U} \cup \mathcal{M}$ such that for every $u \in \mathcal{U}$ and $m \in \mathcal{M}$:

1. $\Phi(u) \in \mathcal{M}$;
2. $\Phi(m) \subseteq \mathcal{U}$;
3. $|\Phi(u)| = 1$;
4. $|\Phi(m)| = L_m$;
5. $m = \Phi(u) \Leftrightarrow u \in \Phi(m)$,

where $|\cdot|$ returns the size of the matching. Conditions 1) and 3) state that each user is matched with an RB, while conditions 2) and 4) imply that each RB is matched with L_m users.

Inspired by the many-to-one housing assignment problem [31], I introduce the notion of swap matching into our many-to-one matching model, and propose a matching algorithm for the joint user-RB association and PA problem [12]. A swapping operation means two users matched with different RBs exchange their matches, while the matching for other users remains the same. The PA is then updated for the two corresponding RBs. Note that how to allocate power to obtain the EE for a given RB will be presented in the following sections, and we assume it is known here. To ensure an improved EE performance, a swapping operation is approved and the matching is updated only when the sum of the EEs for the two RBs involved increases after the swap. Then, to maximize the EE of the considered system, the idea is to continue the swapping operation until no swapping is further approved. Pseudocode for the proposed swapping-based algorithm is given in Algorithm 2.

Note that in the initialization phase, the basic idea is to associate the user to the RB in which it has a large channel gain. This leads to either a higher data rate for the user, or a lower transmit power. Both yield a higher EE. Then, for the swap matching phase, iterations will continue until no swapping operation can be approved in a new round.

4.4.2 Convergence and Complexity

Theorem 4.1. *The proposed joint user-RB association and PA algorithm converges after a finite number of swapping operations.*

Proof: After a number of swapping operations, the structure of matching changes as follows:

$$\Phi_0 \to \Phi_1 \to \Phi_2 \to \cdots, \tag{4.5}$$

where Φ_0 is the initial matching. For swapping operation l, the matching changes from Φ_{l-1} to Φ_l. Denote the corresponding system EE as $\eta_{\mathrm{S}}^{\mathrm{EE}}(l-1)$ and $\eta_{\mathrm{S}}^{\mathrm{EE}}(l)$. Therefore, we have $\eta_{\mathrm{S}}^{\mathrm{EE}}(l) > \eta_{\mathrm{S}}^{\mathrm{EE}}(l-1)$, i.e., the system EE increases at each swapping operation. Moreover, the system EE clearly has an upper bound due to the limited power and spectrum resources. Consequently, the number of potential swapping operations is finite. ■

Given the convergence of the proposed algorithm, we discuss its computational complexity. For the initial phase, it takes $O(U^2M)$ operations. In the swap matching phase, denote the number of iterations to reach the final matching as I_1.[1] In each iteration, all possible swapping combinations should be considered, which requires $O(U^2)$ operations. In each swapping attempt, we need to conduct PA to calculate the EE before and after the swapping for the two related clusters. Denote the computational complexity of the power allocation for calculating the EE as $O(X)$, which will be given in the following section. Then, the total complexity for the swap matching phase is $O(I_1U^2X)$. Adding this to the initial phase, we obtain the total complexity as $O(U^2(I_1X + M))$.

[1]This number cannot be given in closed form, since we do not know for sure at which iteration the proposed algorithm reaches the final matching. This is quite common in the design of most heuristic algorithms. To evaluate the convergence speed, we will show the distribution of this number in the Simulation Results section.

Algorithm 2 Proposed joint user-RB association and PA algorithm.

Step 1: ***Initialization phase***
$K \leftarrow \lceil \frac{U}{M} \rceil$, $\hat{U} \leftarrow U$;
for $k = \{1, \cdots, K\}$
  $\hat{M} \leftarrow M$, Count $\leftarrow 1$;
  while (Count $\leq M$)
    $h_{m^\star,u^\star} \leftarrow \max\{|h_{m,u}|\}|_{\forall m \in \hat{M}, \forall u \in \hat{U}}$
    assign $u^\star$ to RB $m^\star$;
    $\hat{U} \leftarrow \hat{U} \backslash u^\star$, $\hat{M} \leftarrow \hat{M} \backslash m^\star$;
    Count $\leftarrow$ Count $+ 1$;
  end while
end for
Step 2: ***Swap matching phase***
$Indicator = 1$;
while ($Indicator$)
  $Indicator = 0$;
  for $u \in \{1, \cdots, U\}$,
    for $k \in \{1, \cdots, U\}$
      if $\Phi(k) = \Phi(u)$
        continue;
      else
        calculate and compare the EE before and after the swap using Algorithm 2;
        if EE increases
          update the matching, $Indicator \leftarrow 1$;
        end if
      end if
    end for
  end for
end while

4.5 Power Allocation under Given User-RB Association

In line 21 of Algorithm 2, it is assumed that the way of allocating power under a given user-RB association is known. In this section, I present in detail how we conduct PA to maximize the EE. Under a given user-RB association, we can conclude that maximizing the system EE is equivalent to maximizing the EE for each cluster. This is because the system EE is the summation over all cluster EEs, which are mutually independent as they are allocated with different RBs. As a result, we can consider the EE maximization problem on each RB separately, and the mth subproblem is given by

$$\max_{P_{m,l}} \ \eta_m^{\mathrm{EE}} \tag{4.6a}$$

$$\text{s.t.} \ \ R_{m,l} \geq R_{m,l}^{\min}, l \in \{1, \cdots, L_m\} \tag{4.6b}$$

$$P_{m,l} \leq P_{m,l}^{\max}, l \in \{1, \cdots, L_m\}. \tag{4.6c}$$

As the considered subproblems have the same form on different RBs, in the following sections, we omit the RB index m for notational simplicity. Also, L_m is replaced by L while η_m^{EE} is replaced by η_{EE}.

4.5.1 Determine the Feasibility

Owing to the minimum rate requirements and transmit power constraints, (4.6) may be infeasible, i.e., there may not exist a PA solution to satisfy all the constraints. As a result, we need to find the feasibility conditions first. Observe that the last user receives no interference from other users due to SIC; we start with it and obtain

$$\log_2\left(1+\frac{P_L|h_L|^2}{\sigma^2}\right) \geq R_L^{\min} \Leftrightarrow P_L \geq \frac{(2^{R_L^{\min}}-1)}{|h_L|^2}. \tag{4.7}$$

To satisfy the above requirement, we have

$$P_L^{\max} \geq \frac{(2^{R_L^{\min}} - 1)}{|h_L|^2}. \tag{4.8}$$

Assume that (4.8) is satisfied. Clearly, to reduce the interference from user L to other users, it needs to use the minimum transmit power, i.e., $P_L = P_L^{\min} = \frac{(2^{R_L^{\min}} - 1)}{|h_L|^2}$. Now we consider the $(L-1)$th user. Likewise, we have

$$\begin{aligned} &\log_2\left(1 + \frac{P_{L-1}|h_{L-1}|^2}{P_L|h_L|^2 + \sigma^2}\right) \geq R_{L-1}^{\min} \\ &\Leftrightarrow P_{L-1} \geq \frac{2^{R_L^{\min}}(2^{R_{L-1}^{\min}} - 1)}{|h_{L-1}|^2} \\ &\Rightarrow P_{L-1}^{\max} \geq \frac{2^{R_L^{\min}}(2^{R_{L-1}^{\min}} - 1)}{|h_{L-1}|^2}. \end{aligned} \tag{4.9}$$

Using the mathematical induction, we can easily extend it to all users, and obtain

$$P_l^{\max} \geq P_l^{\min} = \frac{2^{\sum_{k=l+1}^{L} R_k^{\min}}(2^{R_l^{\min}} - 1)}{|h_l|^2}, \forall l \in \{1, \cdots, L\}, \tag{4.10}$$

where $P_l^{\min}$ is the minimum power required to satisfy the minimum rate requirement for the lth user. Here we assume that $\sum_{k=L+1}^{L} R_k^{\min} = 0$. Once the above conditions between the minimum rate requirements and the power constraints are satisfied, (4.6) is feasible.

4.5.2 Maximizing the EE when (4.6) is Feasible

The objective function (4.6a) is of fractional form, which is non-convex and challenging to handle. To tackle it, we first deal with the numerator, i.e., the sum rate, which can be

re-written as

$$R^{\text{sum}} = \sum_{l=1}^{L} \log_2 \left(1 + \frac{P_l |h_l|^2}{\sum_{k=l+1}^{L} P_k |h_k|^2 + \sigma^2}\right) = \log_2 \left(1 + \frac{\sum_{l=1}^{L} P_l |h_l|^2}{\sigma^2}\right). \tag{4.11}$$

It is easy to see that the sum rate is a concave function with respect to (w.r.t.) the transmit power for each user.

Now the QoS constraints (4.6b) is considered. It is non-convex on its current form. However, after some mathematical manipulations, it can be reformulated as

$$P_l |h_l|^2 \geq \left(2^{R_l^{\min}} - 1\right) \left(\sum_{k=l+1}^{L} P_k |h_k|^2 + \sigma^2\right), \tag{4.12}$$

which is a linear constraint, since it is just an affine mapping w.r.t., $P_l, l \in \{1, \cdots, L\}$.

Accordingly, problem (4.6) can be re-written as

$$\max_{P_l} \ \frac{\log_2 \left(1 + \frac{\sum_{l=1}^{L} P_l |h_l|^2}{\sigma^2}\right)}{P_f + \sum_{l=1}^{L} P_l} \tag{4.13a}$$

$$\text{s.t.} \ \ (4.12)\ (4.6c), l \in \{1, \cdots, L\}. \tag{4.13b}$$

For the objective function (4.13a), its numerator is a strictly concave function w.r.t., $P_l,\ l \in \{1, \cdots, L\}$, while the denominator is an affine mapping over $P_l,\ l \in \{1, \cdots, L\}$. Therefore, it is a strictly pseudo-concave function [32, Proposition 6]. According to the property of strictly pseudo-concave function, it can be optimally solved by applying the Dinkelbach's algorithm [32, Proposition 6]. The specific procedure is summarized in Algorithm 3. Denote the number of iterations for Algorithm 3 to converge as I_2. During

each iteration, the proposed algorithm needs to solve the following problem, i.e., line 4,

$$\max_{P_l} \log_2\left(1+\frac{\sum_{l=1}^{L} P_l|h_l|^2}{\sigma^2}\right)-\beta(P_f+\sum_{l=1}^{L} P_l) \tag{4.14a}$$

$$\text{s.t. } (4.12)\ (4.6c), l\in\{1,\cdots,L\}, \tag{4.14b}$$

where β is known. Clearly, the above problem is concave, and can be solved using standard algorithms, such as interior-point method. However, the standard approach does not exploit the specific structure of (4.14), and is computationally intensive when (4.14) needs to be solved over and over again. This is indeed the case here, since solving Algorithm 3 requires solving Algorithm 3 many times, i.e., line 21, and addressing Algorithm 3 also requires to solve (4.14) many times, i.e., line 4. To relieve the computational burden, we propose a low-complexity optimal solution for (4.14) as follows:

Denote $F=\log_2\left(1+\frac{\sum_{l=1}^{L} P_l|h_l|^2}{\sigma^2}\right)-\beta(P_f+\sum_{l=1}^{L} P_l)$. Then, for user l, we have $\frac{\partial F}{\partial P_l}=\frac{|h_l|^2}{\ln 2(\sum_{k=1}^{L} P_k|h_k|^2+\sigma^2)}-\beta$. Setting $\frac{\partial F}{\partial P_l}=0$, we obtain $P_l^0=\frac{1}{\beta\ln 2}-\frac{\sum_{k\neq l} P_k|h_k|^2+\sigma^2}{|h_l|^2}$. If all other power values, i.e., $P_k, k\neq l$ are fixed, we can easily obtain the optimal solution for user l by comparing P_l^0 with its minimum and maximum power constraints. Specifically, the optimal power $P_l^\star$ is given by

$$P_l^\star=\begin{cases} P_l^{\min}, & \text{if } P_l^0<P_l^{\min},\\ P_l^{\max}, & \text{if } P_l^0>P_l^{\max},\\ P_l^0, & \text{otherwise.}\end{cases} \tag{4.15}$$

On this basis, the proposed low-complexity algorithm goes as follows: we first allocate the minimum required power to each user; then, we update the power for the users one by one using (4.15); this update continues until convergence. Note that convergence is guaranteed since F increases or remains unchanged after each update, and there exists an

upper bound. Denote the number of iterations for convergence as I_3; then, its complexity is just $O(I_3)$. Thus, we have $X = I_2 I_3$, and $O(U^2(I_1 X + M)) = O(U^2(I_1 I_2 I_3 + M))$. Moreover, the obtained local optimum is also the global optimum since (4.14) is concave. The specific procedure is summarized in Algorithm 4.

Remark. *For any other SIC order, it can be easily proved that the objective function is the same as (4.13a). Moreover, the minimum rate constraints can be turned into convex constraints similar to* (4.12)*. Therefore, Algorithms 2 and 3 can be directly used for EE maximization under any other SIC order.*

Algorithm 3 Proposed EE maximization PA algorithm.

1: **Initialize parameters.**
2: Set $\epsilon > 0; \beta \leftarrow 0; F > \epsilon$;
3: **while** $F > \epsilon$ **do**
4: $P_l^\star \leftarrow \text{argmax } \log_2\left(1 + \frac{\sum_{l=1}^{L} P_l |h_l|^2}{\sigma^2}\right) - \beta(P_f + \sum_{l=1}^{L} P_l)$; s.t. (4.12) (4.6c);
5: $F \leftarrow \log_2\left(1 + \frac{\sum_{l=1}^{L} P_l^\star |h_l|^2}{\sigma^2}\right) - \beta(P_f + \sum_{l=1}^{L} P_l^\star)$;
6: $\beta \leftarrow \frac{\log_2\left(1+\frac{\sum_{l=1}^{L} P_l^\star |h_l|^2}{\sigma^2}\right)}{P_f + \sum_{l=1}^{L} P_l^\star}$;
7: **end while**

Although the proposed Algorithms 3 and 4 can be used to solve the considered EE maximization problem, they do not shed much light into the behaviour of the system, since an iterated algorithm is used. For example, how much power will be employed by the user with the highest channel gain? To this end, several important properties are observed and listed in the sequel:

Lemma 4.1. *Transferring power*[2] *from a user with lower channel gain to a user with higher channel gain leads to increased EE.*

[2]Note that here transferring power means one user lowers his transmit power, while another user increases the same amount of transmit power.

Algorithm 4 Proposed low-complexity algorithm for (4.14).

Initialize parameters.
  Set $P_l \leftarrow P_l^{\min}, l \in \{1, \cdots, L\}$;
while 1 **do**
  $P_{\text{old}} \leftarrow P$;
  for $l = \{1, \cdots, L\}$;
    $P_l^0 = \frac{1}{\beta \ln 2} - \frac{\sum_{k \neq i} P_k |h_k|^2 + \sigma^2}{h_i|^2}$;
    if $P_l^0 < P_l^{\min}$
      $P_l \leftarrow P_l^{\min}$;
    elseif $P_l^0 > P_l^{\max}$
      $P_l \leftarrow P_l^{\max}$;
    else
      $P_l \leftarrow P_l^0$;
    end if
    if $|P_{\text{old}} - P| < 10^{-9}$
      break;
    end if
  end for
end while

Proof: According to (4.13a), when the power transfer happens, the numerator increases owing to the channel gain ordering. On the other hand, the sum transmit power remains unchanged, and thus, the denominator remains unchanged. Therefore, the EE increases as well. ■

Theorem 4.2. *If* $\frac{\partial \eta_{\text{EE}}}{\partial P_1}|_{P_1^{\max}, \bar{P}_{-1}} \geq 0$*, user 1 should transmit at full power to maximize the EE, where* $\bar{P}_{-1} = \bar{P}_2, \cdots, \bar{P}_L$ *denotes a feasible PA solution for the other users.*

Proof: First, when $P_1 = P_1^{\max}$, the feasible region for the other users is maximized, since the interference from user 1 is cancelled by SIC, and the minimum rate requirement of user 1 is most likely to be satisfied. This means that if there exists a feasible region, $P_1 = P_1^{\max}$ is inside it. Furthermore, since $\frac{\partial \eta_{\text{EE}}}{\partial P_1}|_{P_1^{\max},\ \bar{P}_{-1}} \geq 0$, then $P_1 = P_1^{\max}$ maximizes the EE, when $\bar{P}_{-1}$ remains fixed [32, Proposition 5]. Next, we consider the case when power transfer happens between user 1 and other users. According to Lemma 4.1, transferring power from user 1 to other users leads to a lower EE. Therefore, this will not happen.

This completes the proof. ■

Table 4.1: PA Solution for Two Users under Case I.

Phases	**Conditions**	**Solutions**
Phase I	$\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1}\big\vert_{P_1^{\max},P_2^{\max}} \geq \frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}\big\vert_{P_1^{\max},P_2^{\max}} \geq 0$	$P_1 \leftarrow P_1^{\max}$, $P_2 \leftarrow \min\left(P_2^{\max}, \frac{P_1^{\max}\lvert h_1\rvert^2}{(2^{R_1^{\min}}-1)\lvert h_2\rvert^2} - \frac{\sigma^2}{\lvert h_2\rvert^2}\right)$
Phase II	$\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1}\big\vert_{P_1^{\max},P_2^{\max}} \geq 0$, $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}\big\vert_{P_1^{\max},P_2^{\max}} \leq 0$ $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1}\big\vert_{P_1^{\max},P_2^{\min}} \geq 0 \left(P_2^{\min} = \frac{(2^{R_2^{\min}}-1)\sigma^2}{\lvert h_2\rvert^2}\right)$	$P_1 \leftarrow P_1^{\max}$, set $P_2^\star \leftarrow \frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}\big\vert_{P_1^{\max}} = 0$; if $P_2^\star \leq P_2^{\min}$, then $P_2 \leftarrow P_2^{\min}$ else $P_2 \leftarrow \min\left(\frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}\big\vert_{P_1^{\max}} = 0, \frac{P_1^{\max}\lvert h_1\rvert^2}{(2^{R_1^{\min}}-1)\lvert h_2\rvert^2} - \frac{\sigma^2}{\lvert h_2\rvert^2}\right)$;
Phase III	$\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1}\big\vert_{P_1^{\max},P_2^{\max}} \leq 0$, $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}\big\vert_{P_1^{\max},P_2^{\max}} \leq 0$, $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1}\big\vert_{P_1^{\max},P_2^{\min}} \geq 0 \left(P_2^{\min} = \frac{(2^{R_2^{\min}}-1)\sigma^2}{\lvert h_2\rvert^2}\right)$	Same as Phase II
Phase IV	$\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1}\big\vert_{P_1^{\max},P_2^{\min}} \leq 0$, $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}\big\vert_{P_1^{\max},P_2^{\min}} \leq 0$	$P_1 \leftarrow \max\left(\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1}\big\vert_{P_2^{\min}} = 0, \frac{(2^{R_1^{\min}}-1)2^{R_2^{\min}}\sigma^2}{\lvert h_1\rvert^2}\right)$, $P_2 \leftarrow P_2^{\min}$

Table 4.2: PA Solution for Two Users under Case II.

Phases	**Conditions**	**Solutions**
Phase I	$\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1}\big\vert_{P_1^{\max},P_2^{\max}} \geq \frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}\big\vert_{P_1^{\max},P_2^{\max}} \geq 0$	$P_1 \leftarrow \min\left(P_1^{\max}, \frac{P_2^{\max}\lvert h_2\rvert^2}{(2^{R_2^{\min}}-1)\lvert h_1\rvert^2} - \frac{\sigma^2}{\lvert h_1\rvert^2}\right)$, $P_2 \leftarrow P_2^{\max}$
Phase II	$\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1}\big\vert_{P_1^{\max},P_2^{\max}} \geq 0$, $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}\big\vert_{P_1^{\max},P_2^{\max}} \leq 0$	if $\overline{P}_1^\star \leq P_1^{\min}$, $P_1 \leftarrow \overline{P}_1^{\min}$, $P_2 \leftarrow (\overline{P}_1^{\min} - b)/k$ if $\overline{P}_1^\star \in [\overline{P}_1^{\min}, P_1^{\max}]$, $P_1 \leftarrow \overline{P}_1^\star$, $P_2 \leftarrow \overline{P}_2^\star$ if $\overline{P}_1^\star \geq P_1^{\max}$, $P_1 \leftarrow P_1^{\max}$, $P_2 \leftarrow \frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}\big\vert_{P_1^{\max}} = 0$
Phase III	$\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1}\big\vert_{P_1^{\max},P_2^{\max}} \leq 0$, $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}\big\vert_{P_1^{\max},P_2^{\max}} \leq 0$	same as Phase II

Theorem 4.3. *If* $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1}\big\vert_{P_1^{\max},P_{-1}^{\min}} \leq 0$, *then* $P_l = P_l^{\min}, l \neq 1$ *and* $P_1 = \max\left(\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1}\big\vert_{P_{-1}^{\min}} = 0, P_1^{\min}\right)$, *where* $P_{-1}^{\min} = P_2^{\min}, \cdots, P_L^{\min}$ *denotes the minimum required power.*

Proof: The derivative $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_l}$ is given by

$$\frac{\partial \eta_{\mathrm{EE}}}{\partial P_l} = \frac{\lvert h_l\rvert^2}{(\sigma^2 + \sum_{l=1}^{L} P_l\lvert h_l\rvert^2)(P_f + \sum_{l=1}^{L} P_l)\ln 2} - \frac{\log_2\left(1 + \frac{\sum_{l=1}^{L} P_l\lvert h_l\rvert^2}{\sigma^2}\right)}{(P_f + \sum_{l=1}^{L} P_l)^2}. \tag{4.16}$$

Clearly, the derivative $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_l}$ is arranged following the same order of $|h_l|^2$, i.e., $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1} \geq \cdots \geq \frac{\partial \eta_{\mathrm{EE}}}{\partial P_l} \cdots \geq \frac{\partial \eta_{\mathrm{EE}}}{\partial P_L}$. Since $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1}|_{P_1^{\max}, P_{-1}^{\min}} \leq 0$, then $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_l}|_{P_1^{\max}, P_{-1}^{\min}} \leq 0, \forall l \in \{2, \cdots, L\}$. Therefore, all users should reduce their transmit power to increase the EE [32, Proposition 5]. On the other hand, for all users except user 1, they can only reduce their power to the minimum required power. So, we have $P_l = P_l^{\min}, l \neq 1$. Once all the other users' powers are fixed, the EE is maximized at the unique root of $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1}|_{P_{-1}^{\min}} = 0$ or the boundary point, i.e., $P_1^{\min}$. Combining this, we can conclude that $P_1 = \max\left(\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1}|_{P_{-1}^{\min}} = 0, P_1^{\min}\right)$. ■

Remark. *Note that the condition for Theorem 4.2 holds when $P_l^{\max}$ are quite small, while that for Theorem 4.3 holds when $P_l^{\max}$ are quite large. Thus, some good insights for these two extreme cases have been derived. However, for the cases between these two extremes, it is quite complicated to derive analytical results due to the coupling between the QoS requirements and power constraints.*

4.6 Two User Case

Although it is challenging to derive the analytical solution for the general case of multiple users, for the special case of two users, this is possible. Since the derivatives $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_l}$ are arranged as $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1} \geq \cdots \geq \frac{\partial \eta_{\mathrm{EE}}}{\partial P_l} \cdots \geq \frac{\partial \eta_{\mathrm{EE}}}{\partial P_L}$, for the two user case, there are only three cases to consider: $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1} \geq \frac{\partial \eta_{\mathrm{EE}}}{\partial P_2} \geq 0$, $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1} \geq 0 \geq \frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}$ and $0 \geq \frac{\partial \eta_{\mathrm{EE}}}{\partial P_1} \geq \frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}$. On the other hand, under different SIC orders, the feasibility region is different, and thus, different PA solutions are required. The two SIC orders need to be discussed separately. In the following, we first consider the case for the SIC order which decodes user 1 first, and refer to it as Case I. Then, the other case is considered, which is referred to as Case II.

4.6.1 Analytical Solution when User 1 is Decoded First

In this case, the PA solutions for the two users are listed in Table 4.1.

Proof: Refer to Appendix. ■

Remark. *The bisection method can be used to find the root for the equation* $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_l} = 0$*, with the complexity of* $\log_2(P_l^{\max}/\delta)$*, where* δ *denotes the required precision. This is also the dominant computation of obtaining the solution for the EE. According to Table 4.1, when the system is in Phases I, II or III, user 1 always transmits at full power. In Phase IV, user 2 transmits at minimum power. Moreover, from Phase I to Phase IV, we can see how the users react when the maximum allowable transmit power increases. In Phase I, the maximum allowable transmit power is too small, and all the transmit power should be consumed not to violate the QoS constraint. In Phases II and III, user 2 should only transmit with the power which ensures both QoS and maximum EE.*

4.6.2 Analytical Solution when User 2 is Decoded First

In this case, the problem can be formulated as

$$\max_{P_1,P_2} \frac{\log_2\left(1+\frac{P_1|h_1|^2+P_2|h_2|^2}{\sigma^2}\right)}{P_f+P_1+P_2} \tag{4.17a}$$

$$\text{s.t.} \quad P_l \le P_l^{\max}, m \in \{1,2\}, \tag{4.17b}$$

$$\log_2\left(1+\frac{P_2|h_2|^2}{P_1|h_1|^2+\sigma^2}\right) \ge R_2^{\min}, \tag{4.17c}$$

$$\log_2\left(1+\frac{P_1|h_1|^2}{\sigma^2}\right) \ge R_1^{\min}, \tag{4.17d}$$

where (4.17c) and (4.17d) represent the QoS requirements for user 2 and user 1, respectively. Note that (4.17a) is the same as (4.13a) for two users. Indeed, for uplink NOMA, if there exists no QoS constraints, the achievable sum rate under any SIC order is the same, and so is the EE. However, with the QoS constraints, the feasibility region of the power may vary under different SIC orders, and thus, leading to different sum rates and EEs.

The corresponding solution for the above problem is listed in Table 4.2. Note that in this table, we have $\overline{P}_1^{\min} = \frac{(2^{R_1^{\min}}-1)\sigma^2}{|h_1|^2}$. Also, we replace P_1 with P_2 by considering that equality is achieved at the minimum rate of user 2. Thus, we have $P_1 = \frac{P_2|h_2|^2}{(2^{R_2^{\min}}-1)|h_1|^2} - \frac{\sigma^2}{|h_1|^2} = kP_2 + b$, with $k = \frac{|h_2|^2}{(2^{R_2^{\min}}-1)|h_1|^2}$ and $b = -\frac{\sigma^2}{|h_1|^2}$. Then, the multi-variable function of the EE becomes a single variable function over P_2, which is given by

$$f(P_2) = \frac{\log_2\left(1 + \frac{(kP_2+b)|h_1|^2+P_2|h_2|^2}{\sigma^2}\right)}{P_f + kP_2 + P_2 + b}. \tag{4.18}$$

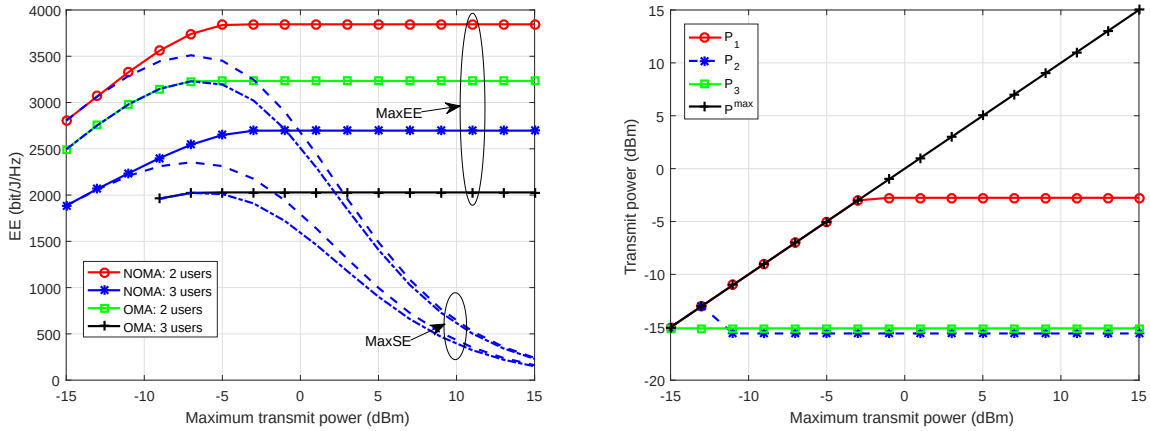
Correspondingly, the root of the derivative is denoted as $\overline{P}_2^{\star} \leftarrow f'(P_2) = 0$. The corresponding value for P_1 is $\overline{P}_1^{\star} = k\overline{P}_2^{\star} + b$.

Proof: Refer to Appendix. ■

Remark. *It can be seen that changing the SIC order leads to different PA results. Under Case II, even for Phases I and II, user 1 may not transmit at full power. For Phase I, instead, user 2 transmits at full power. Also, it is quite difficult to judge which decoding order is better, since this depends on the transmit power constraint and the QoS requirement. If both constraints are the same for both users, under Phase I, it is clear that Case I always outperforms Case II. However, even in this case, except from Phase I, it is still difficult to compare them analytically.*

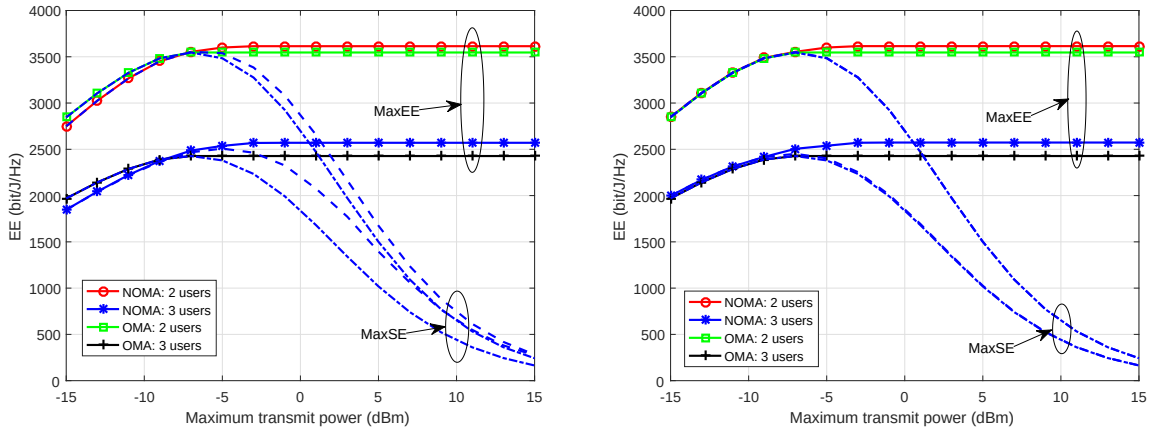
Table 4.3: Simulation Parameters.

Parameters	**Value**
Number of users per cluster	$L = 2, 3$
Number of RBs	$M = 1, 4, 8$
Minimum rate requirement	$R^{\min} = 1.5$ [bit/s/Hz]
Fixed Transmit power per user	$P_f = 0$ [dBm]
Channel bandwidth per RB	180 [KHz]
Noise power spectral density	-174 [dBm/Hz]
Path-loss model	$128 + 35\log_{10}(d)$, d in kilometer
Small scale fading	$\mathcal{CN}(0, 1)$



(a) (b)

Fig. 4.1: Case I: larger channel gain difference; a) EE versus maximum transmit power; b) corresponding transmit power for three users; $|h_1|^2 = 1.10 \times 10^{-9}, |h_2|^2 = 1.34 \times 10^{-10}, |h_3|^2 = 4.25 \times 10^{-11}$.



(a) (b)

Fig. 4.2: Case II: smaller channel gain difference; a) EE versus maximum transmit power with QoS constraints; b) EE versus maximum transmit power without QoS constraints; $|h_1|^2 = 7.31 \times 10^{-10}, |h_2|^2 = 5.81 \times 10^{-10}, |h_3|^2 = 3.10 \times 10^{-10}$.

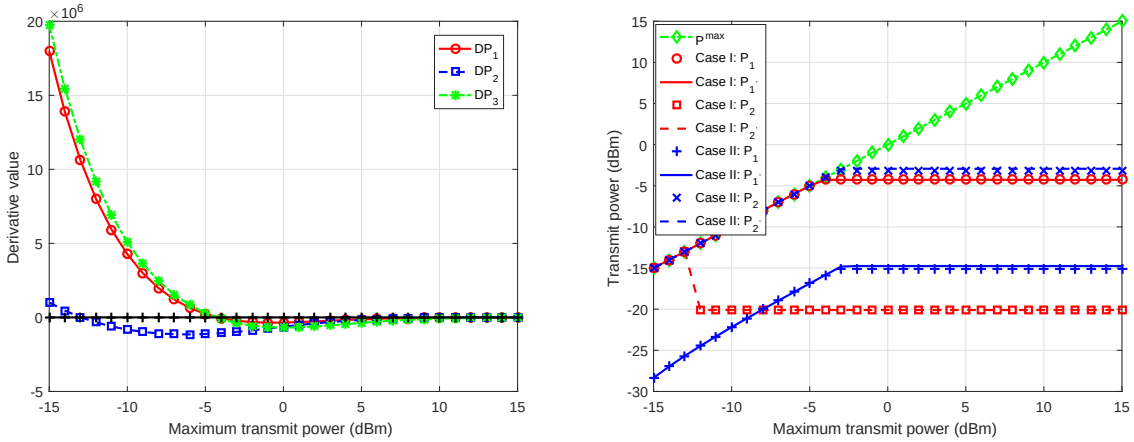
4.7 Simulation Results

In this section, simulations are conducted to verify the developed results. The default simulation parameters are listed in Table 4.3. Note that in simulations, the same minimum rate requirements and maximum transmit power constraints are assigned to all users.

4.7.1 Single Cluster

Results for two cases with different channel gain difference between the users are shown in Figs. 4.1 and 4.2. In addition, as a baseline algorithm, OMA with equal degrees of freedom is presented. The results are also obtained by running the proposed Algorithm 3 with the rate expressions adjusted according to the OMA protocol. Moreover, the results are also presented when the objective is to maximize the SE of the system, which is denoted as "MaxSE". In contrast, the EE maximizing results are denoted as "MaxEE". From Figs. 4.1(a) and 4.2(a), it can be seen that the EE first increases with the maximum transmit power for both "MaxEE" and "MaxSE". Then, after a certain threshold, "MaxEE" saturates, while "MaxSE" continues to decrease. This illustrates the importance of applying an energy-efficient PA algorithm, especially under high maximum transmit power.

Specifically, Fig. 4.1 shows the case with larger channel gain difference among the users. In this case, NOMA achieves much higher EE than OMA for both two and three users, respectively. Moreover, for both NOMA and OMA, the two user case is much better than the three user case. Fig. 4.1(b) shows how the three users allocate their power as the maximum transmit power increases. For user 1, under low maximum transmit power, full power is consumed, which agrees with Theorem 4.2. Under high maximum transmit power, its power no longer increases with the maximum transmit power, but saturates. This coincides with Theorem 4.3. Moreover, for users 2 and 3, they are transmitting using the minimum required power, as expected based on Theorem 4.3.



(a) (b)

Fig. 4: Comparison between two SIC orders; a) Partial derivative values; b) PA; $|h_1|^2 = 1.10 \times 10^{-9}, |h_2|^2 = 1.34 \times 10^{-10}$.

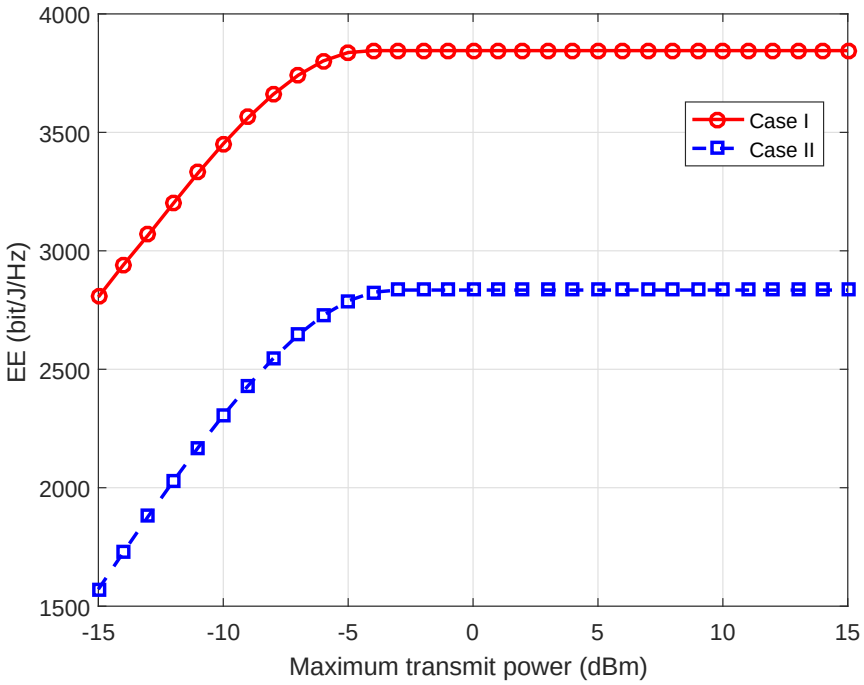


Fig. 3: Comparison of the EE between the two SIC orders; $|h_1|^2 = 1.10 \times 10^{-9}, |h_2|^2 = 1.34 \times 10^{-10}$.

Fig. 4.2 shows the case of smaller channel gain difference. In this case, for both OMA and NOMA, the two user case is better than the three user case. However, under low maximum transmit power, OMA is better than NOMA. This is because the interference introduced by NOMA leads to a smaller feasibility region. Take two user case for example, under low maximum transmit power, OMA is transmitting at full power for both users. However, if $\frac{P_1^{\max}|h_1|^2}{(2^{R_1^{\min}}-1)|h_2|^2} - \frac{\sigma^2}{|h_2|^2} < P_2^{\max}$, user 2 in NOMA cannot transmit at full power to ensure the QoS of user 1. This is quite different from the downlink case, in which the BS controls the PA for all users, and can distribute power among them. In uplink, each user is constrained by its own maximum transmit power. Since user 1's power cannot be increased over its maximum transmit power, the allowable power for user 2 cannot be increased either. Consequently, NOMA achieves lower EE than OMA. Fig. 4.2(b) shows the case when there is no QoS constraints. As expected, NOMA is always better than OMA, even though the gain is quite minor for the two user case. Comparing this with Fig. 4.1, it implies that user pairing should be conducted such that the users' channel gain should be distinct. Moreover, it is worth mentioning that NOMA still outperforms OMA under high maximum transmit power.

Figures 4.3 and 4.4 compare the performance between the two SIC orders. According to Fig. 3, Case I achieves much higher EE than Case II. Fig. 4.4(a) shows how the partial derivative values vary with the maximum transmit power, where DP_1, DP_2 and DP_3 denote $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_1}|_{P_1^{\max},P_2^{\max}}$, $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}|_{P_1^{\max},P_2^{\max}}$ and $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}|_{P_1^{\max},P_2^{\min}}$, respectively. It can be seen that as the maximum transmit power increases, Case I moves from Phase I to Phase IV, while Case II moves from Phase I to Phase III. Fig. 4.4(b) plots P_1 and P_2 obtained by Algorithm 3 (without prime) and the proposed analytical solution (with prime). Obviously, the same results are obtained by both methods, which demonstrates the correctness of the analytical solution.[3] Particularly, under low maximum transmit power, the system is in Phase I,

[3]As we use the log value on the y-axis, it may seem that for Case II, there exists a difference between these two algorithms. Indeed, the difference is smaller than 10^{-5}, and it exists simply because the root

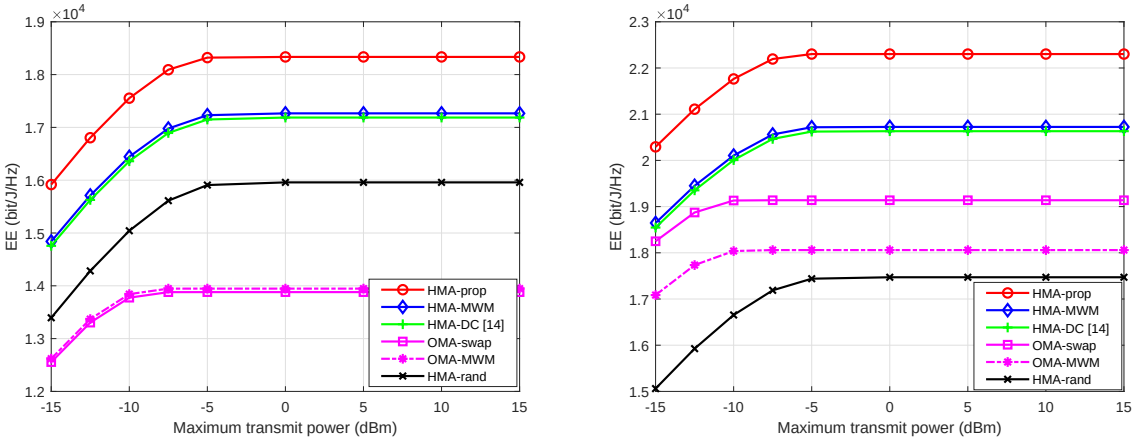
user 1 in Case I and user 2 in Case II transmit at full power.

4.7.2 Multiple Clusters

In this subsection, multiple clusters are considered. The proposed solution, denoted as HMA-prop, is compared with two OMA-based algorithms, i.e., OMA-swap and OMA-MWM, and three HMA-based algorithms, i.e., HMA-DC [14], HMA-MWM and HMA-rand. OMA-swap follows the same procedure as HMA-prop, but with the rate expressions adjusted according to the OMA protocol. In OMA-MWM, we update the PA and user-RB association alternately until convergence. More exactly, under a given PA, it is clear that the EE maximization is equivalent to the sum rate maximization. According to the OMA protocol, the achievable rate of user (m,l) is $R^{\mathrm{O}}_{m,l} = \frac{1}{L_m}\log_2\left(1+\frac{L_m P_{m,l}|h_{m,l}|^2}{\sigma^2}\right)$, which depends only on the allocated RB. Consider the users and RBs as the two set of nodes in a bipartite graph, and the corresponding rates $R^{\mathrm{O}}_{m,l}$ as the weights. Then, the matching that maximizes the sum weight also maximizes the sum rate, and further the EE. This matching can be obtained efficiently using standard maximum weight matching (MWM) algorithms, such as the Hungarian algorithm [33]. Under a given user-RB association, the PA can be solved using Algorithm 2, with the rate expressions adjusted according to the OMA protocol. Note that convergence is guaranteed since the EE increases or remains unchanged after each update, and there exists an upper bound.

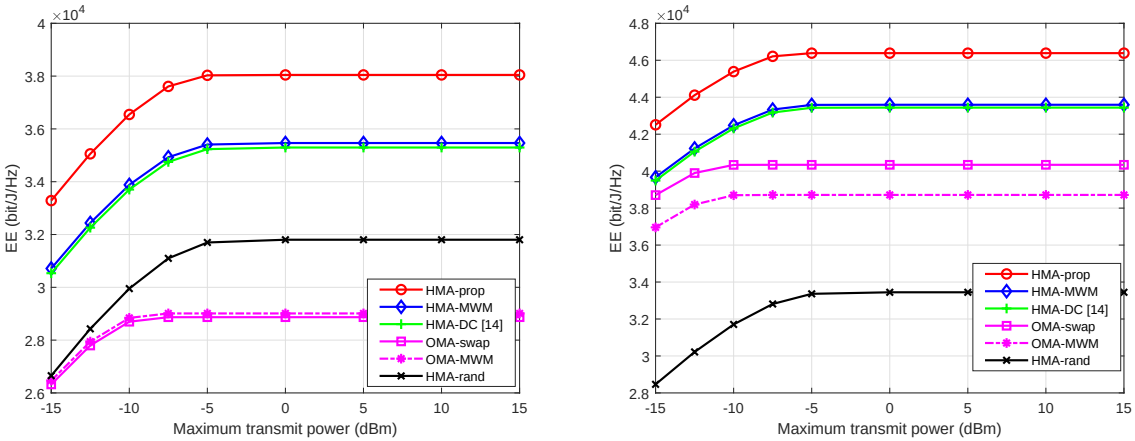
HMA-DC is the scheme proposed in [14], in which each user sends its matching request to its most preferred RB based on the channel gain. However, the preferred RB only accepts the users that lead to the maximum EE. The rejected users will move to the next preferred RB and this process continues until all users are matched to an RB. For HMA-MWM, we cannot apply it the same way as for OMA-MWM, since MWM for HMA cannot be performed due to the intra-cluster interference. Instead, we simply consider the

can only be approximated using the bisection method.



(a) (b)

Fig. 4.5: Comparison of average EE when $U = 12$ and $M = 4$; a) smaller channel gain difference; b) larger channel gain difference.



(a) (b)

Fig. 4.6: Comparison of average EE when $U = 24$ and $M = 8$; a) smaller channel gain difference; b) larger channel gain difference.

weights to be the channel gains. Then, we conduct the MWM between the users and the RBs to achieve the maximum sum channel gains. In HMA-rand, the users are allocated to the RBs randomly.

The following results are averaged over 10^3 random trials, and for each trial, the users' locations are generated following a uniform distribution. The result when there are 12 users accessing 4 RBs is first presented. Two cases with different channel gain differences are considered. Fig. 5(a) shows the result when all users lie within a radius of 150 m. In Fig. 4.5(b), the users are equally divided into three circles, with radii of 50, 100 and 150 m, respectively. Therefore, Fig. 4.5(a) is the case with smaller channel gain difference. It is clear that HMA-prop is the best, followed by HMA-MWM, HMA-DC, HMA-rand OMA-MWM and OMA-swap. This validates the superiority of the proposed scheme over other HMA- and OMA-based algorithms. In Fig. 4.5(b), it can be seen that HMA-prop is still the best, followed by HMA-MWM and HMA-DC. However, in this case, OMA-swap outperforms OMA-MWM, and HMA-rand is the worst. Quite surprisingly, by comparing Figs. 4.5(a) and 4.5(b), it can be observed that a larger channel gain difference does not necessarily lead to a larger gain of HMA over OMA. This does not contradict the conclusion in the single cluster, where we claim that a large channel gain difference among users yields a larger gain of HMA over OMA. This is because the user-RB association results in HMA and OMA can be quite different, and the former conclusion holds when the user-RB association remains the same for both schemes.

Figure 4.6 shows the result when the number of users and RBs are doubled, i.e., now there are 24 users accessing 8 RBs. By comparing Figs. 4.5 and 4.6, it is clear that the corresponding EE values in Fig. 4.6 are more than twice those in Fig. 4.5, except for HMA-rand. This implies that a multiplexing gain is obtained by having more RBs.

Figure 4.7 plots the cumulative distribution function (CDF) of the number of swapping operation required to reach convergence for the above two scenarios. It can be seen

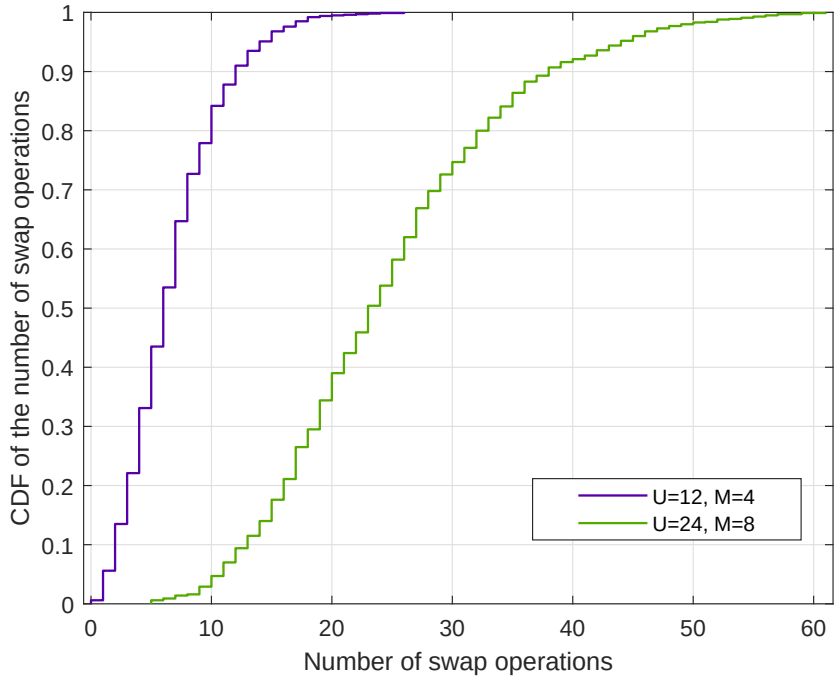


Fig. 4.7: CDF of the number of swap operations for convergence.

that the number of swapping operations grows with that of RBs. However, less than 60 swapping operations are needed even for the scenario when $U = 24$ and $M = 8$, which is quite small. In addition, for Algorithm 3, simulation results show that exactly 6 iterations are required for it to converge when there are 3 users sharing an RB.

4.8 Conclusion

In this chapter, the energy-efficient resource allocation for HMA uplink with QoS requirements was studied for each user. Based on swap matching in many-to-one bipartite graph, a joint user-RB association and power allocation scheme was proposed, which is guaranteed to converge. Under a given user-RB association, it was shown that the system EE maximization equals cluster EE maximization. Then, the feasibility conditions were derived, and the EE maximization was solved using Dinkelbach's algorithm. Moreover, to

further relieve the computational burden, a low-complexity optimal algorithm was proposed for solving the convex optimization subproblem inside the Dinkelbach's algorithm. For the two user case, analytical solutions were further derived for the two SIC orders. Simulations were performed, which verify the developed analytical solutions. Moreover, the results for a single cluster show that under low maximum transmit power, OMA can be better than HMA for uplink, due to the smaller feasibility region for HMA caused by the QoS requirements. On the other hand, under high maximum transmit power, HMA still outperforms OMA. Results under multiple clusters fully validate the superiority of the proposed scheme over other HMA- and OMA-based algorithms. Furthermore, a multiplexing gain can be observed when employing more RBs.

Appendix

Proof of Table I

In Phases I, II and III, as it satisfies the condition for Theorem 4.2, it can be concluded that $P_1 = P_1^{\max}$. As for Phase IV, it is exactly the condition for Theorem 4.3, and thus, the conclusion holds. Then, we only need to prove the PA for user 2 in Phases I, II and III.

Let us first consider Phase I. For the differentiable strictly pseudo-concave function, since $\frac{\partial \eta_{\text{EE}}}{\partial P_1}|_{P_1^{\max},P_2^{\max}} \geq \frac{\partial \eta_{\text{EE}}}{\partial P_2}|_{P_1^{\max},P_2^{\max}} \geq 0$, we can conclude that $\frac{\partial \eta_{\text{EE}}}{\partial P_1} \geq \frac{\partial \eta_{\text{EE}}}{\partial P_2} \geq 0$ for any value of P_1 and P_2. Thus, increasing the transmit power for each user leads to an larger EE. However, for user 2, increasing P_2 also causes more interference to user 1. To ensure the QoS requirement of user 1, the maximum power can be used by user 2 is given by $\frac{P_1^{\max}|h_1|^2}{(2^{R_1^{\min}}-1)|h_2|^2} - \frac{\sigma^2}{|h_2|^2}$. Combining this with the transmit power constraint, we have $P_2 = \min\left(P_2^{\max}, \frac{P_1^{\max}|h_1|^2}{(2^{R_1^{\min}}-1)|h_2|^2} - \frac{\sigma^2}{|h_2|^2}\right)$.

Next, let us focus on Phases II and III. Without considering the QoS constraint, P_2

is obtained when $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}|_{P_1^{\max}} = 0$. Denote it as $P_2^{\star}$, satisfying $P_2^{\star} < P_2^{\max}$. On the other hand, due to the minimum rate requirements for user 1 and user 2, P_2 has a lower bound $P_2^{\min}$, and an upper bound $\frac{P_1^{\max}|h_1|^2}{(2^{R_1^{\min}}-1)|h_2|^2} - \frac{\sigma^2}{|h_2|^2}$. If $P_2^{\star} \leq P_2^{\min}$, $P_2 = P_2^{\min}$. Otherwise, $P_2 = \min\left(P_2^{\star}, \frac{P_1^{\max}|h_1|^2}{(2^{R_1^{\min}}-1)|h_2|^2} - \frac{\sigma^2}{|h_2|^2}\right)$.

Proof of Table II

In Phase I, due to the change of SIC order, user 2 should transmit at full power. As for user 1, it should not violate the QoS requirement of user 2, and thus, $P_1 < \frac{P_2^{\max}|h_2|^2}{(2^{R_2^{\min}}-1)|h_1|^2} - \frac{\sigma^2}{|h_1|^2}$. Combining it with the maximum power, we have $P_1 = \min\left(P_1^{\max}, \frac{P_2^{\max}|h_2|^2}{(2^{R_2^{\min}}-1)|h_1|^2} - \frac{\sigma^2}{|h_1|^2}\right)$.

In Phases II and III, it is first assumed that P_1 is constrained by the QoS of user 2. Then, we turn the multi-variable function into a single variable function. Accordingly, the solution for this is the root of the derivative. Denote this as $\overline{P}_2^{\star}$. Correspondingly, $\overline{P}_1^{\star} = k\overline{P}_2^{\star} + b$. On the other hand, user 1 needs to satisfy its own QoS, and thus, it can be obtained that $\overline{P}_1^{\min}$. If $\overline{P}_1^{\star} < \overline{P}_1^{\min}$, $P1 = \overline{P}_1^{\min}$, and $P_2 = (\overline{P}_1^{\min} - b)/k$. If $P_2 < (\overline{P}_1^{\min} - b)/k$, it cannot satisfy its own QoS. If P_2 exceeds this, EE decreases, since $P_2 > \overline{P}_2^{\star}$, and $P_1 > \overline{P}_1^{\star}$. When $\overline{P}_1^{\star}$ lies in $(\overline{P}_1^{\min}, P_1^{\max})$, if $\overline{P}_2^{\star} \leq P_2^{\max}$, $P_1 = \overline{P}_1^{\star}$, $P_2 = \overline{P}_2^{\star}$ is clearly the solution. As for the case $\overline{P}_2^{\star} > P_2^{\max}$, this cannot hold. This is because we can transfer power from user 2 to user 1, and increase the EE. Therefore, $\overline{P}_2^{\star}$ cannot be the root of (4.18). When $\overline{P}_1^{\star} > P_1^{\max}$, $P_1 = P_1^{\max}$. Denote the root of $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}|_{P_1^{\max}}$ as P_2^{r}. Since $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}|_{P_1^{\max}, P_2^{\max}} \leq 0$, $P_2^{\mathrm{r}} \leq P_2^{\max}$. In addition, we can obtain $P_2^{\mathrm{r}} \geq (P_1^{\max} - b)/k$, owing to $\frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}|_{P_1^{\max}, (P_1^{\max}-b)/k} \geq \frac{\partial \eta_{\mathrm{EE}}}{\partial P_2}|_{\overline{P}_1^{\star}, \overline{P}_2^{\star}} = 0$. Therefore, the QoS of user 2 can be satisfied when

$P_1 = P_1^{\max}$ and $P_2 = P_2^{\mathrm{r}}$. In sum, it can be concluded that $P_2 = P_2^{\mathrm{r}}$.

References

[1] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation for uplink NOMA," in *Proc IEEE Globecom*, Abu Dhabi, UAE, Dec. 2018, pp. 1–6.

[2] S. M. R. Islam, M. Zeng, O. A. Dobre, and K. Kwak, "Resource allocation for downlink NOMA systems: Key techniques and open issues," *IEEE Wireless Commun. Mag.*, vol. 25, no. 2, pp. 40–47, Apr. 2018.

[3] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[4] S. M. R. Islam, M. Zeng, and O. A. Dobre, "NOMA in 5G systems: Exciting possibilities for enhancing spectral efficiency," *IEEE 5G Tech. Focus,* vol. 1, no. 2, May 2017. [Online]. Available: 307 http://5g.ieee.org/tech-focus.

[5] L. Dai, B. Wang, Y. Yuan, S. Han, C. L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[6] S. M. R. Islam, et al., "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surv. Tuts.*, vol. 19, no. 2, pp. 721–742, Second quarter 2017.

[7] F. Zhou, Y. Wu, R. Q. Hu, Y. Wang, and K. K. Wong, "Energy-efficient NOMA enabled heterogeneous cloud radio access networks," *IEEE Network*, vol. 32, no. 2, pp. 152–160, Mar. 2018.

[8] M. Zeng, G. I. Tsiropoulos, O. A. Dobre, and M. H. Ahmed, "Power allocation for cognitive radio networks employing non-orthogonal multiple access," in *Proc. IEEE Global Commun. Conf.*, Washington DC, USA, Dec. 2016.

[9] 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, Dec. 2014.

[10] 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, Aug. 2017.

[11] ——, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Select. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.

[12] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686–7698, Nov. 2016.

[13] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "A fair individual rate comparison between MIMO-NOMA and MIMO-OMA," in *Proc IEEE Globecom Wkshps*, Singapore, Dec 2017, pp. 1–5.

[14] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722–3732, Sep. 2016.

[15] Y. Zhang, H. M. Wang, T. X. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2852–2857, Mar. 2017.

[16] W. M. Hao et al., "Energy-efficient power allocation in millimeter wave massive MIMO with non-orthogonal multiple access," *IEEE Wireless Commun. Lett.*, vol. PP, no. 99, pp. 1–1, Jun. 2017.

[17] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation for MIMO-NOMA with multiple users in a cluster," *IEEE Access*, vol. 6, pp. 5170–5181, Feb. 2018.

[18] W. Hao, Z. Chu, F. Zhou, S. Yang, G. Sun, and K. Wong, "Green communication for NOMA-based CRAN," *IEEE Internet of Things J.*, pp. 1–1, 2018.

[19] T. Lv, Y. Ma, J. Zeng, and P. T. Mathiopoulos, "Millimeter-wave NOMA transmission in cellular M2M communications for internet of things," *IEEE Internet of Things J.*, vol. 5, no. 3, pp. 1989–2000, Jun. 2018.

[20] Y. Endo, Y. Kishiyama, and K. Higuchi, "Uplink non-orthogonal access with MMSE-SIC in the presence of inter-cell interference," in *Proc IEEE ISWCS*, Aug. 2012, pp. 261–265.

[21] X. Chen, A. Benjebbour, A. Li, and A. Harada, "Multi-user proportional fair scheduling for uplink non-orthogonal multiple access (NOMA)," in *Proc. IEEE VTC*, May. 2014, pp. 1–5.

[22] W. Liu, X. Hou, and L. Chen, "Enhanced uplink non-orthogonal multiple access for 5G and beyond systems," *Front. Inform. Technol. Electron. Eng.*, vol. 19, no. 3, pp. 340–356, Mar. 2018.

[23] Y. Liang, X. Li, J. Zhang, and Z. Ding, "Non-orthogonal random access for 5G networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4817–4831, July 2017.

[24] M. Shirvanimoghaddam, M. Condoluci, M. Dohler, and S. J. Johnson, "On the fundamental limits of random non-orthogonal multiple access in cellular massive IoT," *IEEE J. Select. Areas Commun.*, vol. 35, no. 10, pp. 2238–2252, Oct. 2017.

[25] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X. Xia, "Joint power control and beamforming for uplink non-orthogonal multiple access in 5g millimeter-wave communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6177–6189, Sep. 2018.

[26] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (noma) systems," *IEEE Access*, vol. 4, pp. 6325–6343, Aug. 2016.

[27] D. Zhai and J. Du, "Spectrum efficient resource management for multi-carrier-based NOMA networks: A graph-based method," *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 388–391, Jun. 2018.

[28] T. Lv, Z. Lin, P. Huang, and J. Zeng, "Optimization of the energy-efficient relay-based massive IoT network," *IEEE Internet of Things J.*, vol. 5, no. 4, pp. 3043–3058, Aug. 2018.

[29] Z. Yang, W. Xu, H. Xu, J. Shi, and M. Chen, "Energy efficient non-orthogonal multiple access for machine-to-machine communications," *IEEE Commun. Lett.*, vol. 21, no. 4, pp. 817–820, Apr. 2017.

[30] M. Zeng, W. Hao, O. A. Dobre, and V. Poor, "Energy-efficient power allocation in uplink mmwave massive MIMO with NOMA," *IEEE Trans. Veh. Technol.*, pp. 1–1, 2019.

[31] E. Bodine-Baron, C. Lee, A. Chong, B. Hassibi, and A. Wierman, "Peer effects and stability in matching markets," in *Proc. 4th Symp. Algorithmic Game Theory (SAGT)*, Amalfi, Italy, Oct. 2011, pp. 117–129.

[32] A. Zappone, P. Lin, and E. Jorswieck, "Energy efficiency in secure multi-antenna systems," *IEEE Trans. Signal Process.,* submitted for publication. [Online]. Available: http://arxiv.org/abs/1505.02385.

[33] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics (NRL)*, vol. 1-2, pp. 83–97, 1955.

Chapter 5

Securing Downlink Massive MIMO-NOMA Networks with Artificial Noise

5.1 Abstract

This chapter focuses on securing the confidential information of massive MIMO-NOMA networks by exploiting artificial noise (AN). An uplink training scheme is first proposed with minimum mean squared error estimation at the base station. Based on the estimated channel state information, the base station precodes the confidential information and injects the AN. Following this, the ergodic secrecy rate is derived for downlink transmission. An asymptotic secrecy performance analysis is also carried out for a large number of transmit antennas and high transmit power at the base station, respectively, to highlight the effects of key parameters on the secrecy performance of the considered system. Based on the derived ergodic secrecy rate, the joint power allocation of the uplink training phase and downlink transmission phase is proposed to maximize the sum secrecy rates

of the system. Besides, from the perspective of security, another optimization algorithm is proposed to maximize the energy efficiency. The results show that the combination of massive MIMO technique and AN greatly benefits NOMA networks in term of the secrecy performance. In addition, the effects of the uplink training phase and clustering process on the secrecy performance are revealed. Besides, the proposed optimization algorithms are compared with other baseline algorithms through simulations, and their superiority is validated. Finally, it is shown that the proposed system outperforms the conventional massive MIMO orthogonal multiple access in terms of the secrecy performance.

5.2 Introduction

The development of Internet-of-Things demands massive connectivity over the limited radio spectrum. This requires the next generation wireless networks deploy new multiple access technologies with better spectral efficiency [1]. Recently, NOMA has been introduced as a solution for this challenge [2, 3]. Power-domain NOMA allows multiple users to share the same time-frequency resource simultaneously by using superposition coding and advanced interference cancellation techniques, such as SIC [4–6]. As a result, NOMA can enhance the capacity of a network in both spatial and temporal dimensions [7–10]. However, from the security viewpoint, sharing the same time-frequency resource among users imposes secrecy challenges.

Traditionally, the security issues have been handled at the higher layers using encryption approaches. However, the development of computing technologies and the tremendous growth in the number of wireless devices have surfaced the vulnerability of the conventional encryption methods [11]. As a result, physical layer security (PLS) has been introduced as an additional protecting layer to the conventional encryption methods for securing confidential information [12]. The principle of PLS is to take advantage of the

randomness of the wireless channels to restrain the illegitimate side from overhearing the legitimate users [13]. The community has shown a great interest in applying PLS to NOMA networks. In [14], the authors investigated the secrecy outage probability (SOP) of NOMA relay networks with two types of relay, i.e., amplify-and-forward and decode-and-forward. The chapter revealed that in the high signal-to-noise ratio regime, the SOP of the considered NOMA relay network converges to a constant value. In [15], the secrecy performance of a stochastic NOMA network was considered, by modelling its users' locations using stochastic geometry. The results showed that the secrecy diversity order of the considered system is determined by that of the user pair with a poorer channel. In [16], the authors derived a closed-form solution for maximizing the secrecy sum rate of the NOMA while taking the users' quality of service requirements into consideration. In [17], the authors investigated a NOMA system in the presence of an external eavesdropper. The SOP of the considered system was derived and used to optimize the decoding order, transmission rates, and allocated power. These studies have laid the initial foundation for exploiting PLS in NOMA networks.

Recently, massive MIMO has become one of the key technologies for 5G network [18–20]. By deploying hundreds of antennas at the BS to serve tens of users, massive MIMO exploits the high spatial resolution and large array gain to greatly enhance the throughput, SE, and EE [21–23]. Massive MIMO networks are suggested to operate in time division duplex to address pilot contamination by exploiting channel reciprocity [18]. In massive MIMO networks, the BS can obtain the knowledge of the CSI via uplink training sequences of the users and employ this knowledge to precode the transmit data. The combination of massive MIMO and NOMA seems to be naturally matched since it can offer a great performance enhancement for a large number of users [24]. However, there are some challenges of this combination. Since the number of orthogonal sequences for the uplink training phase is limited, the massive number of users has to be grouped

in clusters. In a cluster, users share the same training sequence. As a consequence, the quality of the uplink training phase can be compromised. Therefore, the spatial resolution is decreased, which can lead to leakage of the confidential information. There have been several studies of PLS for massive MIMO-NOMA networks. In [25], the authors have investigated the secrecy performance of a NOMA massive MIMO network in the presence of an active eavesdropper. The inter-user interference was utilized to enhance the secrecy performance of the network. Artificial noise (AN) has proven its effectiveness to secure the legitimate side from malicious attempts [26, 27]. Recently, in [28], the authors have proposed a joint alignment of multi-user constellations and AN to secure the massive MIMO-NOMA networks. Therefore, the role of AN in massive MIMO-NOMA networks is far from being well-understood.

In this chapter, an AN-based PLS method is proposed for the massive MIMO-NOMA networks in the presence of a passive eavesdropper. In order to secure the downlink transmission, the BS uses its knowledge of CSI to precode the confidential information and inject the AN, which is different from [25]. Besides, because of the high complexity of the uplink training phase in the massive MIMO-NOMA networks, the AN approaches in [26, 27] are not suitable. Therefore, the AN is injected in the null-space of the effective channels of the clusters in the downlink transmission phase. To emphasize the role of the uplink training process on the secrecy performance of the considered system, the CSI knowledge at the BS is the result of an estimation process that is more practical than the assumption of perfect CSI in other existing work on PLS for massive MIMO-NOMA networks. To the best of our knowledge, this is the first work using AN to secure massive MIMO-NOMA networks when taking imperfect channel estimation into account. The contributions of this chapter can be summarized as follows:

- This chapter demonstrates a framework to analyze the secrecy performance of an AN-aided massive MIMO-NOMA network while taking the imperfect channel es-

timation into consideration. In particular, the ergodic secrecy rates for users are derived. The asymptotic expressions of the legitimate and illegitimate rates for a large number of antennas and high transmit power at the BS are also obtained. Note that the AN-aided massive MIMO-OMA network is a special case of the proposed system. The analysis expressions can be applied directly with the number of users in each cluster being equal to one.

- The results reveal that by using a sufficiently large number of antennas at the BS, the AN only affects the eavesdropper. In addition, when the transmit power at the BS is sufficiently high, the secrecy performance of a user depends on the AN, the intra-cluster interference, and the channel estimation error of its cluster.

- In order to further exploit the interference and AN, this chapter studies the maximization of the sum ergodic secrecy rate (SE) and the maximization of the EE in terms of the ergodic secrecy rates. In this chapter, the EE is defined as the sum ergodic secrecy rate over the total transmit power, which includes both the uplink and downlink powers. The SE maximization problem is first decomposed into two sub-problems, i.e., uplink and downlink PA, based on alternating optimization. Then, each sub-problem is addressed using difference of convex (DC) programming. The EE maximization problem is of fractional form, and can be transformed into a series of SE maximization problems, which can be solved accordingly. Numerical results show that the proposed algorithms can significantly enhance the performance of the considered system, compared with other baseline algorithms.

The rest of this chapter is organized as follows. The system and channel models are described in Section 5.2. The analytical expressions for the ergodic secrecy rates of the considered system are developed in Section 5.3. In Section 5.4, the optimization problems are proposed, and the solutions are discussed in Section 5.5. The numerical results and

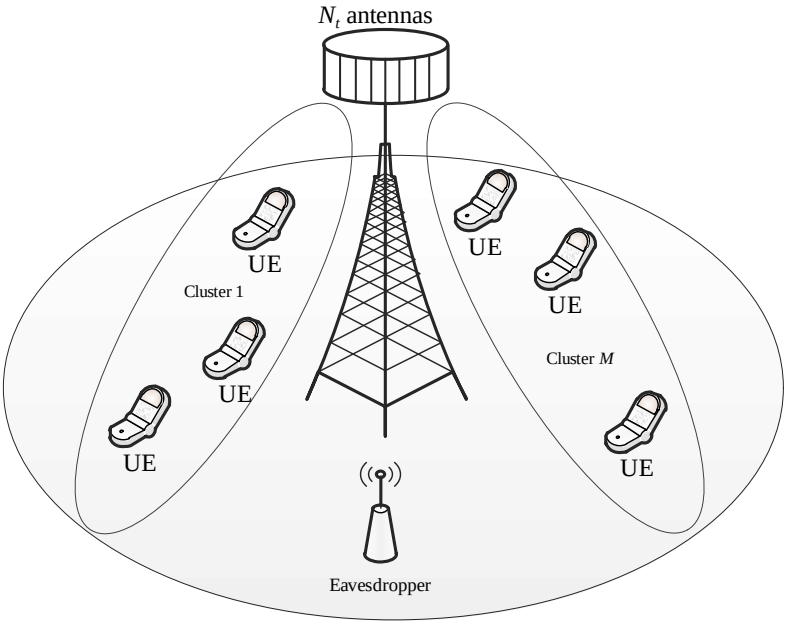


Fig. 5.1: System model.

discussions are presented in Section 5.6. Finally, Section 5.7 concludes the chapter.

Notations

Superscript $(\cdot)^H$ stands for the conjugate transpose. The expectation operation and Frobenius norm are denoted by $\mathbb{E}\{\cdot\}$ and $\|\cdot\|$, respectively. I_{N_t} denotes the N_t-dimensional identity matrix. $\mathcal{CN}(\mu, \sigma^2)$ indicates complex normal distribution with μ mean and σ^2 variance.

5.3 System and Channel Models

As shown in Fig. 5.1, we consider the downlink transmission in a massive MIMO-NOMA system, which includes one N_t-antenna BS, multiple single-antenna end users (UEs) that are grouped into M clusters with K_m users, $m = \{1, .., M\}$, in the m-th cluster, and one passive single-antenna eavesdropper. Before performing the downlink transmission, the BS needs the network's CSI to precode the information and inject the AN. Besides,

the users also require knowledge of the precoding to decode the confidential information. Therefore, the BS and users learn CSI and precoding knowledge in the training phases.

5.3.1 Training Phases

Uplink training

During one coherence interval duration of T samples, the users simultaneously send training sequences to the BS. Users in the same cluster employ the same training sequence. In order to prevent the training sequence of each cluster from interfering with each other, all clusters are assigned mutually orthogonal training sequences of length τ samples, where $T \geq \tau \geq M$. The j-th cluster training sequence is denoted by a $\tau \times 1$ vector $\boldsymbol{\Phi}_j$, where $\boldsymbol{\Phi}_j^H \boldsymbol{\Phi}_i = 0, \forall i \neq j$, $\boldsymbol{\Phi}_j^H \boldsymbol{\Phi}_j = 1$. The received training signal at the BS is

$$\boldsymbol{Y} = \sum_{m=1}^{M} \sum_{k=1}^{K_m} \sqrt{P_{m,k} \beta_{m,k} \tau} \boldsymbol{h}_{m,k} \boldsymbol{\Phi}_m^H + \boldsymbol{N}, \tag{5.1}$$

where $P_{m,k}$ is the transmit power of the k-th UE of the m-th cluster, $\beta_{m,k}$ is the large-scale fading, $\boldsymbol{h}_{m,k}$ is the small-scale fading, $\boldsymbol{h}_{m,k} \sim \mathcal{CN}(0, I_{N_t})$, and the elements of $\boldsymbol{N} \sim \mathcal{CN}(0,1)$ represent the additive white Gaussian noise (AWGN). Since $\boldsymbol{\Phi}_m$ is known at the BS, the BS pre-processes the received signal as follows:

$$\begin{aligned} \underbrace{\boldsymbol{Y} \boldsymbol{\Phi}_m}_{\tilde{\boldsymbol{y}}_m} &= \sum_{k=1}^{K_m} \sqrt{P_{m,k} \beta_{m,k} \tau} \boldsymbol{h}_{m,k} + \underbrace{\boldsymbol{N} \boldsymbol{\Phi}_m}_{\tilde{\boldsymbol{n}}_m} \\ &= \sqrt{\sum_{k=1}^{K_m} P_{m,k} \beta_{m,k} \tau} \boldsymbol{h}_m + \tilde{\boldsymbol{n}}_m, \end{aligned} \tag{5.2}$$

where $\boldsymbol{h}_m = \frac{\sum_{k=1}^{K_m} \sqrt{P_{m,k} \beta_{m,k} \tau} \boldsymbol{h}_{m,k}}{\sqrt{\sum_{k=1}^{K_m} P_{m,k} \beta_{m,k} \tau}}$ is the effective channel for the m-th cluster.

The BS uses the minimum mean squared error (MMSE) technique to estimate $\boldsymbol{h}_m$.[1]

[1]The use of MMSE has been widely addopted in massive MIMO system [19, 20].

The estimate of $\boldsymbol{h}_m$ is [25]

$$\hat{\boldsymbol{h}}_m = \frac{\sqrt{\sum_{k=1}^{K_m} P_{m,k}\beta_{m,k}\tau}}{1+\sum_{k=1}^{K_m} P_{m,k}\beta_{m,k}\tau}\tilde{\boldsymbol{y}}_m. \tag{5.3}$$

The relation between $\boldsymbol{h}_{m,k}$ and $\hat{\boldsymbol{h}}_m$ is

$$\boldsymbol{h}_{m,k} = \sqrt{\rho_{m,k}}\hat{\boldsymbol{h}}_m + \sqrt{1-\rho_{m,k}}\boldsymbol{\varepsilon}_{m,k}, \tag{5.4}$$

where $\boldsymbol{\varepsilon}_{m,k} \sim \mathcal{CN}(0, \boldsymbol{I}_{N_t})$ is the error vector, which is independent of $\hat{\boldsymbol{h}}_m$. Besides, $\rho_{m,k} = \frac{P_{m,k}\beta_{m,k}\tau}{1+\sum_{i=1}^{K_m} P_{m,i}\beta_{m,i}\tau}$ [25].

Remark. *For each cluster, the error of the estimation process depends on the uplink transmit power of each user, the number of users in a cluster, the large-scale fading, and the length of the training sequences. This error can be reduced by decreasing the number of users in a cluster. However, this leads to an increase in the number of clusters, and further yields more orthogonal training sequences, which are limited in certain cases, e.g., crowded stadium, busy city center, etc.*

After the estimation process, the BS uses the estimates of the cluster's effective channels to precode. In this chapter, we assume that the BS employs the maximal ratio transmission (MRT) precoder, which is simple and nearly optimal in massive MIMO networks [20]. The precoder is defined as

$$\boldsymbol{w}_m = \frac{\hat{\boldsymbol{h}}_m}{\left\|\hat{\boldsymbol{h}}_m\right\|}. \tag{5.5}$$

Downlink training

The downlink training phase is similar to the uplink training phase, except that the BS uses the obtained precoder to beam the downlink pilots to the clusters. Since the

downlink pilots are known at the users, these users can estimate accurately their effective channel gains, i.e., $|\sqrt{\beta_{m,k}}\boldsymbol{h}_{m,k}^H\boldsymbol{w}_m|^2$. We assume that the estimation process at users is perfect.[2] Without loss of generality, the users' effective channel gains of the m-th cluster are ordered as follows:

$$|\sqrt{\beta_{m,1}}\boldsymbol{h}_{m,1}^H\boldsymbol{w}_m|^2 \geq \cdots \geq |\sqrt{\beta_{m,K_m}}\boldsymbol{h}_{m,K_m}^H\boldsymbol{w}_m|^2. \tag{5.6}$$

During this phase, the eavesdropper also obtains its effective channel gain, i.e., $|\sqrt{\beta_{\mathsf{E}}}\boldsymbol{g}^H\boldsymbol{w}_m|^2$, where β_{E} is the large-scale fading and $\boldsymbol{g}$ is the small-scale fading vector corresponding to the eavesdropper.

5.3.2 NOMA Downlink Transmission

In order to perform NOMA downlink transmission, the BS conducts superposition coding for each cluster. The superposition coding for the m-th cluster is as follows:

$$x_m = \sum_{k=1}^{K_m} \sqrt{Q_{m,k}}s_{m,k}, \tag{5.7}$$

where $Q_{m,k}$ is the transmit power allocated to $\text{UE}_{m,k}$, and $s_{m,k}$ is the corresponding transmitted signal, satisfying $\mathbb{E}\{|s_{m,k}|^2\} = 1$. For securing the confidential information, the BS injects AN into the transmitted signals. The BS combines all cluster signals as follows:

$$\boldsymbol{x} = \sum_{m=1}^{M} (\boldsymbol{w}_m x_m + \sqrt{Q_{m,0}}\boldsymbol{z}_m\lambda_m), \tag{5.8}$$

[2]This assumption is reasonable since it has been proven that at a sufficiently high transmit power, the error of the channel estimation process at the receiver is sufficiently small and can be neglected [27].

where $\boldsymbol{w}_m$ and $\boldsymbol{z}_m$ are the precoding vector and AN vector for the m-th cluster, respectively, $\hat{\boldsymbol{h}}_m^H \boldsymbol{z}_m = 0$, $\|\boldsymbol{z}_m\|^2 = 1$; $Q_{m,0}$ is the power allocated for the AN and λ_m is the AN signal of the m-th cluster, $\mathbb{E}\{|\lambda_m|^2\} = 1$.

The received signal at the $\mathrm{UE}_{m,k}$ is

$$\begin{aligned} y_{m,k} = & \underbrace{\sqrt{\beta_{m,k}} \boldsymbol{h}_{m,k}^H \sqrt{Q_{m,k}} \boldsymbol{w}_m s_{m,k}}_{\text{Desired signal}} \\ & + \underbrace{\sqrt{\beta_{m,k}} \boldsymbol{h}_{m,k}^H \left(\sum_{i=1, i \neq k}^{K_m} \sqrt{Q_{m,i}} \boldsymbol{w}_m s_{m,i} + \sqrt{Q_{m,0}} \boldsymbol{z}_m \lambda_m \right)}_{\text{Intra-cluster interference and AN}} \\ & + \underbrace{\sqrt{\beta_{m,k}} \boldsymbol{h}_{m,k}^H \sum_{j=1, j \neq m}^{M} \left(\sum_{i=1}^{K_j} \sqrt{Q_{j,i}} \boldsymbol{w}_j s_{j,i} + \sqrt{Q_{j,0}} \boldsymbol{z}_j \lambda_j \right)}_{\text{Inter-cluster interference and AN}} \\ & + n_{m,k}, \end{aligned} \tag{5.9}$$

where $n_{m,k} \sim \mathcal{CN}(0,1)$ is the AWGN at $\mathrm{UE}_{m,k}$.

The eavesdropper tries to intercept the confidential information of $\mathrm{UE}_{m,k}$. The received signal at the eavesdropper is

$$\begin{aligned} y_{m,k}^e = & \underbrace{\sqrt{\beta_{\mathsf{E}}} \boldsymbol{g}^H \sqrt{Q_{m,k}} \boldsymbol{w}_m s_{m,k}}_{\text{Desired signal}} \\ & + \underbrace{\sqrt{\beta_{\mathsf{E}}} \boldsymbol{g}^H \left(\sum_{i=1, i \neq k}^{K_m} \sqrt{Q_{m,i}} \boldsymbol{w}_m s_{m,i} + \sqrt{Q_{m,0}} \boldsymbol{z}_m \lambda_m \right)}_{\text{Intra-cluster interference and AN}} \\ & + \underbrace{\sqrt{\beta_{\mathsf{E}}} \boldsymbol{g}^H \sum_{j=1, j \neq m}^{M} \left(\sum_{i=1}^{K_j} \sqrt{Q_{j,i}} \boldsymbol{w}_j s_{j,i} + \sqrt{Q_{j,0}} \boldsymbol{z}_j \lambda_j \right)}_{\text{Inter-cluster interference and AN}} \\ & + n_e, \end{aligned} \tag{5.10}$$

where $n_e \sim \mathcal{CN}(0,1)$ is the AWGN at the eavesdropper.

5.4 Secrecy Performance Analysis

In this section, we derive the ergodic secrecy rate of UE $_{m,k}$ from its ergodic legitimate rate and its corresponding ergodic eavesdropping rate.

5.4.1 Ergodic Secrecy Rate

The ergodic secrecy rate of $\text{UE}_{m,k}$ is

$$\begin{aligned} R_{m,k}^{sec} &= \mathbb{E}\left\{[R_{m,k} - R_{m,k}^{e}]^{+}\right\} \\ &\approx \left[\mathbb{E}\left\{R_{m,k}\right\} - \mathbb{E}\left\{R_{m,k}^{e}\right\}\right]^{+}, \end{aligned} \tag{5.11}$$

where $[x]^{+} = \max(x, 0)$. This approximation is reasonable in massive MIMO systems owing to the channel hardening property [29]. The achievable rate of $\text{UE}_{m,k}$ is

$$\bar{R}_{m,k} = \mathbb{E}\left\{R_{m,k}\right\} \approx \left(1 - \frac{\tau}{T}\right)\log_2(1 + \bar{\gamma}_{m,k}), \tag{5.12}$$

where $\mathbb{E}\{\cdot\}$ denotes the expectation operator and $\bar{\gamma}_{m,k} = \frac{\kappa_{m,k}}{\sum_{t=1}^{3} \Im_{m,k,t} + 1}$, with

$$\begin{aligned} \kappa_{m,k} &= \left|\mathbb{E}\left\{\sqrt{Q_{m,k}\beta_{m,k}}\boldsymbol{h}_{m,k}^{H}\boldsymbol{w}_m\right\}\right|^2 \\ &= Q_{m,k}\beta_{m,k}\left|\mathbb{E}\left\{(\sqrt{\rho_{m,k}}\hat{\boldsymbol{h}}_m^{H}\boldsymbol{w}_m + \sqrt{1-\rho_{m,k}}\boldsymbol{\varepsilon}_{m,k}^{H}\boldsymbol{w}_m)\right\}\right|^2 \\ &\overset{(a)}{=} Q_{m,k}\beta_{m,k}\rho_{m,k}\left|\mathbb{E}\left\{\left\|\hat{\boldsymbol{h}}_m\right\|\right\}\right|^2 \\ &\overset{(b)}{=} Q_{m,k}\beta_{m,k}\rho_{m,k}\frac{\Gamma^2\left(N_t + \frac{1}{2}\right)}{\Gamma^2(N_t)} \\ &\overset{(c)}{\approx} Q_{m,k}\beta_{m,k}\rho_{m,k}N_t, \end{aligned} \tag{5.13}$$

where step (a) holds true because $\mathbb{E}\left\{\boldsymbol{\varepsilon}_{m,k}^H \boldsymbol{w}_m\right\} = \mathbb{E}\left\{\boldsymbol{\varepsilon}_{m,k}^H\right\}\mathbb{E}\left\{\boldsymbol{w}_m\right\} = 0$, $\Gamma(\cdot)$ is the gamma function, step (b) is based on the fact that $\left\|\hat{\boldsymbol{h}}_m\right\|$ has a scaled Chi distribution with $2N_t$ degrees of freedom by a factor of $\frac{1}{\sqrt{2}}$ [27]. Therefore, $\mathbb{E}\left\{\left\|\hat{\boldsymbol{h}}_m\right\|\right\} = \frac{\Gamma\left(N_t+\frac{1}{2}\right)}{\Gamma(N_t)}$, and step (c) is obtained by using the approximation $\frac{\Gamma^2\left(N_t+\frac{1}{2}\right)}{\Gamma^2(N_t)} \overset{N_t\to\infty}{\to} N_t$ [30].

Further, $\Im_{m,k,i}$ for $i = \{1,2,3\}$ in the expression of $\bar{\gamma}_{m,k}$ are given in (5.14), (5.15) and (5.16), respectively. Note that step (a) in (5.15) is obtained because $\hat{\boldsymbol{h}}_m^H \boldsymbol{z}_m = 0$ and $\boldsymbol{\varepsilon}_{m,k}$ is independent of $\boldsymbol{z}_m$.

$$
\begin{aligned}
\Im_{m,k,1} &= Q_{m,k}\beta_{m,k}\left(\mathbb{E}\left\{|\boldsymbol{h}_{m,k}^H \boldsymbol{w}_m|^2\right\} - \left(\mathbb{E}\left\{\boldsymbol{h}_{m,k}^H \boldsymbol{w}_m\right\}\right)^2\right) \\
&= Q_{m,k}\beta_{m,k}\left(\mathbb{E}\left\{\left|\sqrt{\rho_{m,k}}\hat{\boldsymbol{h}}_m^H \boldsymbol{w}_m + \sqrt{1-\rho_{m,k}}\boldsymbol{\varepsilon}_{m,k}^H \boldsymbol{w}_m\right|^2\right\} - \left(\mathbb{E}\left\{\boldsymbol{h}_{m,k}^H \boldsymbol{w}_m\right\}\right)^2\right) \\
&= Q_{m,k}\beta_{m,k}\left(\rho_{m,k}\mathbb{E}\left\{\left|\hat{\boldsymbol{h}}_m^H \boldsymbol{w}_m\right|^2\right\} + (1-\rho_{m,k})\mathbb{E}\left\{\left|\boldsymbol{\varepsilon}_{m,k}^H \boldsymbol{w}_m\right|^2\right\} - \left(\mathbb{E}\left\{\boldsymbol{h}_{m,k}^H \boldsymbol{w}_m\right\}\right)^2\right) \\
&= Q_{m,k}\beta_{m,k}\left(\rho_{m,k}N_t + 1 - \rho_{m,k} - \rho_{m,k}\frac{\Gamma^2\left(N_t+\frac{1}{2}\right)}{\Gamma^2(N_t)}\right) \\
&= Q_{m,k}\beta_{m,k}(1-\rho_{m,k}),
\end{aligned} \tag{5.14}
$$

$$
\begin{aligned}
\Im_{m,k,2} &= \mathbb{E}\left\{\beta_{m,k}\left(\sum_{i=1}^{k-1} Q_{m,i}|\boldsymbol{h}_{m,k}^H \boldsymbol{w}_m|^2 + Q_{m,0}|\boldsymbol{h}_{m,k}^H \boldsymbol{z}_m|^2\right)\right\} \\
&= \beta_{m,k}\left(\sum_{i=1}^{k-1} Q_{m,i}\mathbb{E}\left\{|\boldsymbol{h}_{m,k}^H \boldsymbol{w}_m|^2\right\} + Q_{m,0}\mathbb{E}\left\{|\boldsymbol{h}_{m,k}^H \boldsymbol{z}_m|^2\right\}\right) \\
&\overset{(a)}{=} \beta_{m,k}\left[\sum_{i=1}^{k-1} Q_{m,i}(\rho_{m,k}N_t + 1 - \rho_{m,k}) + Q_{m,0}(1-\rho_{m,k})\right],
\end{aligned} \tag{5.15}
$$

$$\Im_{m,k,3} = \mathbb{E}\left\{\beta_{m,k}\sum_{j=1,j\neq m}^{M}\left(\sum_{i=1}^{K_j} Q_{j,i}|\boldsymbol{h}_{m,k}^H\boldsymbol{w}_j|^2 + Q_{j,0}|\boldsymbol{h}_{m,k}^H\boldsymbol{z}_j|^2\right)\right\}$$
$$= \beta_{m,k}\sum_{j=1,j\neq m}^{M}\sum_{i=0}^{K_j} Q_{j,i}. \tag{5.16}$$

It can be seen that $\Im_{m,k,1}$ denotes the desired signal leakage due to the imperfect uplink channel estimation, while $\Im_{m,k,2}$ represents the intra-cluster interference after SIC and the AN leakage. In addition, $\Im_{m,k,3}$ expresses the inter-cluster interference and AN.

Remark. *Note that perfect SIC is assumed to obtain $\Im_{m,k,2}$. That is, the k-th user first decodes and subtracts the interfering signals from the K_m-th to the $(k+1)$-th user in sequence, and then demodulates its desired signal $s_{m,k}$. In other words, the residual intra-cluster interference is only from the users with stronger channel gains, i.e., the first user to the $(k-1)$-th user. In practice, owing to channel estimation error, hardware limitation, low signal quality, and so on, the decoding error of the weak interfering signal may occur. Consequently, there exists residual interference from the weak users after SIC, namely imperfect SIC. This residual interference is similar to the intra-cluster interference. As shown in [31–33], the residual interference can be modeled as a linear function of the power of the interfering signal, and the coefficient of imperfect SIC can be obtained through long-term measurements. As a result, the ergodic secrecy rate in the presence of imperfect SIC can be directly derived by adding the term of residual interference in $\Im_{m,k,2}$.*

The ergodic eavesdropping rate corresponding to $\mathrm{UE}_{m,k}$ is

$$\bar{R}_{m,k}^e = \mathbb{E}\left\{R_{m,k}^e\right\} \approx \left(1-\frac{\tau}{T}\right)\log_2(1+\bar{\gamma}_{m,k}^e), \tag{5.17}$$

where $\bar{\gamma}_{m,k}^e = \frac{\kappa_{m,k}^e}{\sum_{t=1}^{2}\Im_{m,k,t}^e+1}$, with

$$
\begin{aligned}
\kappa^{e}_{m,k} &= Q_{m,k}\beta_{\mathsf{E}}\mathbb{E}\left\{|\boldsymbol{g}^{H}\boldsymbol{w}_{m}|^{2}\right\} = Q_{m,k}\beta_{\mathsf{E}},\\
\Im^{e}_{m,k,1} &= \sum_{i=1,i\neq k}^{K_m} Q_{m,i}\beta_{\mathsf{E}}\mathbb{E}\left\{|\boldsymbol{g}^{H}\boldsymbol{w}_{m}|^{2}\right\}\\
&\quad + Q_{m,0}\beta_{\mathsf{E}}\mathbb{E}\left\{\left|\boldsymbol{g}^{H}\boldsymbol{z}_{m}\right|^{2}\right\}\\
&= \beta_{\mathsf{E}}\sum_{i=0,i\neq k}^{K_m} Q_{m,i},\\
\Im^{e}_{m,k,2} &= \beta_{\mathsf{E}}\sum_{j=1,j\neq m}^{M}\left(\sum_{i=1}^{K_j} Q_{j,i}\mathbb{E}\left\{|\boldsymbol{g}^{H}\boldsymbol{w}_{j}|^{2}\right\}\right.\\
&\quad \left. + Q_{j,0}\mathbb{E}\left\{\left|\boldsymbol{g}^{H}\boldsymbol{z}_{m}\right|^{2}\right\}\right)\\
&= \beta_{\mathsf{E}}\sum_{j=1,j\neq m}^{M}\sum_{i=0}^{K_j} Q_{j,i}.
\end{aligned}
$$

Therefore, $\bar{R}^{e}_{m,k}$ can be simplified as (5.18) as follows:[3]

$$
\bar{R}^{e}_{m,k} = \left(1-\frac{\tau}{T}\right)\log_2\left(1+\frac{Q_{m,k}\beta_{\mathsf{E}}}{\beta_{\mathsf{E}}\sum_{i=0,i\neq k}^{K_m} Q_{m,i} + \beta_{\mathsf{E}}\sum_{j=1,j\neq m}^{M}\sum_{i=0}^{K_j} Q_{j,i} + 1}\right). \tag{5.18}
$$

By comparing the intra-cluster interference terms in $\bar{R}_{m,k}$ and $\bar{R}^{e}_{m,k}$, i.e., $\Im_{m,k,2}$ and $\Im^{e}_{m,k,1}$, we can observe that the intra-cluster interference has less impact on the legitimate users owing to SIC. This helps to achieve a higher secrecy rate.

5.4.2 Asymptotic Secrecy Performance

In this subsection, increasing the number of antennas and the transmit power at the BS are respectively studied to reveal insights into the considered system.

[3]It is possible to extend this work to the case of multiple eavesdroppers or multi-antenna eavesdropper since (5.18) can be applied to each eavesdropper or each antenna of a multi-antenna eavesdropper. The secrecy performance in these cases is determined by the strongest eavesdropper or the strongest eavesdropping antenna.

Large Number of Antennas at the BS

We first investigate the impact of a large number of antennas at the BS on the secrecy performance. From (5.18), we can observe that the eavesdropping rate is independent of the number of antennas at the BS. When this number is large, the legitimate rate is expressed as

$$\bar{R}_{m,k} \overset{N_t \to \infty}{=} \left(1 - \frac{\tau}{T}\right) \log_2 \left(1 + \frac{Q_{m,k}}{\sum\limits_{i=1}^{k-1} Q_{m,i}}\right). \tag{5.19}$$

Remark. *When the number of antennas at the BS is sufficiently large, the secrecy rate converges to a constant value. At the legitimate side, the effect of imperfect CSI, fading, inter-cluster interference, and AN leakage is negligible because of channel hardening. The legitimate rate depends only on the intra-cluster transmit powers. Meanwhile, the eavesdropping rate suffers from noise, interferences, and fading. Obviously, by using AN, the secrecy performance can be guaranteed in this scenario.*

High Transmit Power at the BS

In order to reveal the impact of the transmit power at the BS, the transmit power for each user is set proportional to the maximum transmit power of the BS, i.e., $Q_{m,k} = \sigma_{m,k} Q_{\max}$, where $Q_{\max}$ is the maximum transmit power at the BS and $\sum\limits_{m=1}^{M} \sum\limits_{k=1}^{K_m} \sigma_{m,k} = 1$. When $Q_{\max}$ is large, the legitimate rate and the eavesdropping rate are respectively approximated as (5.20) and (5.21)

$$\bar{R}_{m,k} \overset{Q_{\max} \to \infty}{=} \left(1 - \frac{\tau}{T}\right) \tag{5.20}$$

$$\times \log_2 \left(1 + \frac{\sigma_{m,k} \rho_{m,k} N_t}{\sigma_{m,k}(1 - \rho_{m,k}) + \left[\sum\limits_{i=1}^{k-1} \sigma_{m,i}(\rho_{m,k} N_t + 1 - \rho_{m,k}) + \sigma_{m,0}(1 - \rho_{m,k})\right] + \sum\limits_{j=1, j \neq m}^{M} \sum\limits_{i=0}^{K_j} \sigma_{i,j}}\right),$$

$$\bar{R}^{e}_{m,k} \overset{Q_{\max}\to\infty}{=} \left(1-\frac{\tau}{T}\right)\log_2\left(1+\frac{\sigma_{m,k}}{\sum_{i=0,i\neq k}^{K_m}\sigma_{m,i}+\sum_{j=1,j\neq m}^{M}\sum_{i=0}^{K_j}\sigma_{i,j}}\right). \tag{5.21}$$

Remark. *When the transmit power at the BS is high, we can observe that:*

- *The secrecy rate converges to a constant value. This value is independent of fading and the maximum transmit power.*
- *The legitimate rate and the eavesdropping rate suffer from the same amount of inter-cluster interference and inter-cluster AN. In other words, the secrecy rate is independent of the inter-cluster interference and inter-cluster AN.*
- *The eavesdropper is affected by the AN more heavily than the legitimate user. This effect depends on the uplink training process. Recalling Remark 1, we can conclude that the secrecy performance depends on the number of available orthogonal pilots.*

5.5 Optimization Problems

In this section, we consider the optimization of the uplink and downlink PA to fully exploit the potential of the proposed secure massive MIMO-NOMA network. Two system level criteria are respectively considered, i.e., the SE maximization and the EE maximization.

5.5.1 SE Maximization

First, we aim to maximize the SE for the considered system, which is formulated as

$$\max_{\mathbf{P},\mathbf{Q}} \sum_{m=1}^{M}\sum_{k=1}^{K_m} R_{m,k}^{sec} \tag{5.22a}$$

$$\text{s.t. } 0 \le P_{m,k} \le P_{m,k}^{\max}, m \in \{1,\cdots,M\}, \tag{5.22b}$$

$$k \in \{1,\cdots,K_m\},$$

$$Q_{m,k} \ge 0, m \in \{1,\cdots,M\}, k \in \{0,\cdots,K_m\}, \tag{5.22c}$$

$$\sum_{m=1}^{M}\sum_{k=0}^{K_m} Q_{m,k} \le Q_{\max}, \tag{5.22d}$$

where $\mathbf{P} \in \mathcal{R}^{M\times K_m}$ and $\mathbf{Q} \in \mathcal{R}^{M\times(K_m+1)}$ denote the matrix for the uplink and downlink power, respectively. Equations (5.22b) and (5.22d) represent the maximum transmit power constraint for each user in uplink and the total power constraint in downlink, respectively. Note that there exists a one-to-one mapping between $P_{m,k}$ and $\rho_{m,k}$.

5.5.2 EE Maximization

We also consider maximization of EE, defined as the sum ergodic secrecy rate over the total transmit power, which includes both the uplink and downlink power [34–36]. Moreover, for uplink and downlink power, both fixed circuit power and dynamic transmit power are considered [37, 38]. We denote the overall circuit power of the system as P_f. Then, the EE is given as

$$\eta_{\text{EE}} = \frac{\sum_{m=1}^{M}\sum_{k=1}^{K_m} R_{m,k}^{sec}}{\sum_{m=1}^{M}\sum_{k=1}^{K_m} P_{m,k} + \sum_{m=1}^{M}\sum_{k=0}^{K_m} Q_{m,k} + P_f}. \tag{5.23}$$

Accordingly, the EE optimization problem can be expressed as

$$\max_{\mathbf{P},\mathbf{Q}} \eta_{\text{EE}}, \text{ s.t. (5.22b)} - \text{(5.22d)}. \tag{5.24}$$

5.6 Proposed Solutions

5.6.1 SE Maximization

Problem (5.22) is clearly non-convex, owing to the non-convex objective function. Moreover, it can be seen that the uplink power $\mathbf{P}$ and downlink power $\mathbf{Q}$ are coupled in the objective function. This coupling makes (5.22) difficult to handle. To address it, we propose to decompose the original problem into the following two sub-problems:

Uplink Power Allocation for Channel Estimation

For this sub-problem, we assume that the downlink power is appropriately allocated to the users and the AN, i.e., $\mathbf{Q}$ is known and given. Then, the original problem can be simplified as

$$\max_{\mathbf{P}} \sum_{m=1}^{M} \sum_{k=1}^{K_m} R_{m,k}^{sec}, \text{ s.t. (5.22b).} \tag{5.25}$$

Downlink Power Allocation for Data Transmission

Likewise, here we assume that the uplink power is appropriately allocated to the users, i.e., $\mathbf{P}$ is known and given. Then, the original problem is re-expressed as

$$\max_{\mathbf{P}} \sum_{m=1}^{M} \sum_{k=1}^{K_m} R_{m,k}^{sec}, \text{ s.t. (5.22c), (5.22d).} \tag{5.26}$$

For sub-problem (1), since $\mathbf{Q}$ is given, it can be seen that $\bar{R}_{m,k}^{e}$ is a constant. Then, we only need to consider $\bar{R}_{m,k}$. After some mathematical manipulations, $\bar{R}_{m,k}$ can be

expressed as

$$\bar{R}_{m,k} = (1-\frac{\tau}{T})\log_2\left(1+\frac{\kappa_{m,k}}{\sum\limits_{t=1}^{3}\Im_{m,k,t}+1}\right) \tag{5.27}$$
$$= (1-\frac{\tau}{T})\times\log_2\left(1+\frac{a_1\beta_{m,k}\tau P_{m,k}}{a_2\beta_{m,k}\tau P_{m,k}+a_3\tau\sum_{i=1}^{K_m}\beta_{m,i}P_{m,i}+a_3}\right),$$

where $a_1 = Q_{m,k}\beta_{m,k}N_t$, $a_2 = \beta_{m,k}[(N_t-1)\sum_{i=1}^{k-1}Q_{m,i} - Q_{m,0} - Q_{m,k}]$, $a_3 = \beta_{m,k}\sum_{i=0}^{k}Q_{m,i} + \beta_{m,k}\sum_{j\neq m}\sum_{i=0}^{K_j}Q_{j,i} + 1$.

On this basis, we further transform $f = \sum_{m=1}^{M}\sum_{k=1}^{K_m}\bar{R}_{m,k}$ as

$$f = (1-\frac{\tau}{T})\sum_{m=1}^{M}\sum_{k=1}^{K_m}\underbrace{\log_2\left((a_1+a_2)\beta_{m,k}\tau P_{m,k}+a_3\tau\sum_{i=1}^{K_m}\beta_{m,i}P_{m,i}+a_3\right)}_{f_1(\mathbf{P})}$$
$$-(1-\frac{\tau}{T})\sum_{m=1}^{M}\sum_{k=1}^{K_m}\underbrace{\log_2\left(a_2\beta_{m,k}\tau P_{m,k}+a_3\tau\sum_{i=1}^{K_m}\beta_{m,i}P_{m,i}+a_3\right)}_{f_2(\mathbf{P})}. \tag{5.28}$$

Note that $(1-\frac{\tau}{T})$ is a constant, which does not affect the solution and can be removed. Then, (5.25) can be re-expressed as

$$\max_{\mathbf{P}} \quad \sum_{m=1}^{M}\sum_{k=1}^{K_m} f_1(\mathbf{P}) - f_2(\mathbf{P}), \text{ s.t. (5.22b)}, \tag{5.29}$$

where both functions $f_1(\mathbf{P})$ and $f_2(\mathbf{P})$ are concave. Thus, the objective $\sum_{m=1}^{M}\sum_{k=1}^{K_m} f_1(\mathbf{P}) - f_2(\mathbf{P})$ is a DC function. The gradient of f_2 at $P_{j,i}, \forall j \in \{1,\cdots,M\}, i \in \{1,\cdots,K_j\}$ is

given by

$$\nabla f_2(P_{j,i}) = \begin{cases} \frac{(a_2+a_3)\beta_{m,k}\tau/\ln 2}{a_2\beta_{m,k}\tau P_{m,k}+a_3\tau\sum_{i=1}^{K_m}\beta_{m,i}P_{m,i}+a_3}, & j=m, i=k, \\ \frac{a_3\beta_{m,i}\tau/\ln 2}{a_2\beta_{m,k}\tau P_{m,k}+a_3\tau\sum_{i=1}^{K_m}\beta_{m,i}P_{m,i}+a_3}, & j=m, i\neq k, \\ 0, & j\neq m. \end{cases}$$

Algorithm 5 Proposed Power Allocation Algorithm for Sum Rate Maximization.

Initialize $l \leftarrow 0, \varepsilon^\star \leftarrow 1, \varepsilon \leftarrow 10^{-3}$; Initialize feasible downlink power $\mathbf{Q}^{(0)}$;
while $\varepsilon^\star \geq \varepsilon$
  Uplink power allocation:
  while $|\mathbf{P}^{(l)} - \mathbf{P}^{(l-1)}| > 10^{-3}$
    $\mathbf{P}^{(l)} \leftarrow \max \sum_{m=1}^{M}\sum_{k=1}^{K_m}[f_1(\mathbf{P}) - f_2(\mathbf{P}^{(l-1)}) - (P_{m,k} - P_{m,k}^{(l-1)}) \times \sum_{j=1}^{M}\sum_{i=1}^{K_j}\nabla f_2(P_{j,i}{}^{(l-1)})]$ s.t. (5.22b)
  end while
  Downlink power allocation:
  while $|\mathbf{Q}^{(l)} - \mathbf{Q}^{(l-1)}| > 10^{-3}$
    $\mathbf{Q}^{(l)} \leftarrow \max \sum_{m=1}^{M}\sum_{k=0}^{K_m}[g_1(\mathbf{Q}) + g_3(\mathbf{Q}) - g_2(\mathbf{Q}^{(l-1)}) - g_4(\mathbf{Q}^{(l-1)}) - (Q_{m,k} - Q_{m,k}^{(l-1)}) \times \sum_{j=1}^{M}\sum_{i=0}^{K_j}\nabla g_2(Q_{j,i}{}^{(l-1)}) + \nabla g_4(Q_{j,i}{}^{(l-1)})]$
    s.t. (5.22c), (5.22d)
  end while
  $R_{\text{sum}}^d \leftarrow \sum_{m=1}^{M}\sum_{k=1}^{K_m} R_{m,k}^{sec}$;
  $\varepsilon^\star \leftarrow R_{\text{sum}}^d - R_{\text{sum}}^u$
  $l \leftarrow l+1$;
end while
if $R_{m,k}^{sec} < 0, \forall m \in \{1,\cdots,M\}, k \in \{1,\cdots,K_m\}$
  $R_{m,k}^{sec} \leftarrow 0$;
end
$R_{\text{sum}} \leftarrow \sum_{m=1}^{M}\sum_{k=1}^{K_m} R_{m,k}^{sec}$.

The following procedure generates a sequence $\{\mathbf{P}^{(l)}\}$ of improved feasible solutions [39, 40]. Initialized from a feasible $\{\mathbf{P}^{(0)}\}$, $\{\mathbf{P}^{(l)}\}$ is obtained as the optimal solution of

the following convex problem at the l-th iteration:

$$\max_{\mathbf{P}} \quad \sum_{m=1}^{M}\sum_{k=1}^{K_m}\left[f_1(\mathbf{P}) - f_2(\mathbf{P}^{(l-1)}) - \ (P_{m,k} - P_{m,k}^{(l-1)}) \times \sum_{j=1}^{M}\sum_{i=1}^{K_j} \nabla f_2(P_{j,i}{}^{(l-1)})\right]$$
$$\text{s.t. (5.22b).} \tag{5.30}$$

Note that (5.30) can be efficiently solved by available convex software packages [41]. Moreover, since there exists no inter-cluster interference, the sum rate maximization can be done in parallel for each cluster, i.e., the system sum rate maximization equals to the cluster sum rate maximization.

After solving the above problem, we can obtain the value for $\mathbf{P}$. Accordingly, we can obtain $\rho_{m,k}$. On this basis, for sub-problem (2), after some mathematical manipulations, $\bar{R}_{m,k}$ can be expressed as follows:

$$\begin{aligned}\bar{R}_{m,k} &= (1-\frac{\tau}{T})\log_2\left(1+\frac{b_1 Q_{m,k}}{b_2 Q_{m,k} + b_3\sum_{i=1}^{k-1} Q_{m,i} + b_2 Q_{m,0} + \beta_{m,k}\sum_{j\neq m}\sum_{i=0}^{K_j} Q_{j,i} + 1}\right) \\ &= (1-\frac{\tau}{T})\underbrace{\log_2\left((b_1+b_2)Q_{m,k} + b_3\sum_{i=1}^{k-1} Q_{m,i} + b_2 Q_{m,0} + \beta_{m,k}\sum_{j\neq m}\sum_{i=0}^{K_j} Q_{j,i} + 1\right)}_{g_1(\mathbf{Q})} \\ &\quad - (1-\frac{\tau}{T})\underbrace{\log_2\left(b_2 Q_{m,k} + b_3\sum_{i=1}^{k-1} Q_{m,i} + b_2 Q_{m,0} + \beta_{m,k}\sum_{j\neq m}\sum_{i=0}^{K_j} Q_{j,i} + 1\right)}_{g_2(\mathbf{Q})},\end{aligned} \tag{5.31}$$

where $b_1 = \rho_{m,k}\beta_{m,k}N_t$, $b_2 = \beta_{m,k}(1-\rho_{m,k})$, and $b_3 = \beta_{m,k}(\rho_{m,k}N_t + 1 - \rho_{m,k})$.

The gradient of g_2 at $Q_{j,i}, \forall j \in \{1, \cdots, M\}, i \in \{0, \cdots, K_j\}$ is given by

$$\nabla g_2(Q_{j,i}) = \begin{cases} \frac{b_2}{b_2 Q_{m,k} + b_3 \sum_{i=1}^{k-1} Q_{m,i} + b_2 Q_{m,0} + \beta_{m,k} \sum_{j \neq m} \sum_{i=0}^{K_j} Q_{j,i} + 1}, & j = m, i = k \text{ or } 0, \\ \frac{b_3}{b_2 Q_{m,k} + b_3 \sum_{i=1}^{k-1} Q_{m,i} + b_2 Q_{m,0} + \beta_{m,k} \sum_{j \neq m} \sum_{i=0}^{K_j} Q_{j,i} + 1}, & j = m, i = 1, \cdots, k-1, \\ \beta_{m,k}, & j \neq m, \\ 0, & \text{otherwise.} \end{cases} \tag{5.32}$$

Next, let us consider $-\bar{R}^e_{m,k}$, which can be re-written as

$$\begin{aligned} -\bar{R}^e_{m,k} = (1 - \frac{\tau}{T}) \underbrace{\log_2 \left(\beta_E \sum_{i \neq k} Q_{m,i} + \beta_E \sum_{j \neq m} \sum_{i=0}^{K_j} Q_{j,i} + 1 \right)}_{g_3(\mathbf{Q})} \\ - (1 - \frac{\tau}{T}) \underbrace{\log_2 \left(\beta_E \sum_{i=0}^{K_m} Q_{m,i} + \beta_E \sum_{j \neq m} \sum_{i=0}^{K_j} Q_{j,i} + 1 \right)}_{g_4(\mathbf{Q})}. \end{aligned} \tag{5.33}$$

The gradient of g_4 at $Q_{j,i}, \forall j \in \{1, \cdots, M\}, i \in \{0, \cdots, K_j\}$ is given by

$$\nabla g_4(Q_{j,i}) = \frac{\beta_E / \ln 2}{\beta_E \sum_{i=0}^{K_m} Q_{m,i} + \beta_E \sum_{j \neq m} \sum_{i=0}^{K_j} Q_{j,i} + 1}. \tag{5.34}$$

The following procedure generates a sequence $\{\mathbf{Q}^{(l)}\}$ of improved feasible solutions [39, 40]. Initialized from a feasible $\{\mathbf{Q}^{(0)}\}$, $\{\mathbf{Q}^{(l)}\}$ is obtained as the optimal solution of the following convex problem at the l-th iteration:

$$\begin{aligned} &\max_{\mathbf{Q}} \sum_{m=1}^{M} \sum_{k=0}^{K_m} [g_1(\mathbf{Q}) + g_3(\mathbf{Q}) - g_2(\mathbf{Q}^{(l-1)}) - g_4(\mathbf{Q}^{(l-1)}) - \\ &(Q_{m,k} - Q^{(l-1)}_{m,k}) \times \sum_{j=1}^{M} \sum_{i=0}^{K_j} \nabla g_2(Q_{j,i}{}^{(l-1)}) + \nabla g_4(Q_{j,i}{}^{(l-1)})] \\ &\text{s.t. (5.22c), (5.22d).} \end{aligned} \tag{5.35}$$

Note that (5.35) can also be efficiently solved by available convex software packages [41].

Now we have solved the two sub-problems. We repeat them after each other until convergence. Then, for those users with negative rates, we set their rates to zero following the $[\cdot]^+$ operation. The specific procedure is summarized in Algorithm 5.

5.6.2 EE Maximization

It is clear that (5.24) belongs to a fractional problem, which can be transformed into a series of parametric subtractive-form subproblems as (5.36) based on Dinkelbach algorithm [42].

$$\max_{\mathbf{P},\mathbf{Q}} \sum_{m=1}^{M}\sum_{k=1}^{K_m} R_{m,k}^{sec} - \lambda^{(l-1)}\left(\sum_{m=1}^{M}\sum_{k=1}^{K_m} P_{m,k} + \sum_{m=1}^{M}\sum_{k=0}^{K_m} Q_{m,k} + P_f\right), \text{ s.t. (5.22b)} - \text{(5.22d)}. \tag{5.36}$$

Note in (5.36), $\lambda^{(l-1)}$ is a non-negative parameter. Starting from $\lambda^{(0)} = 0$, $\lambda^{(l)}$ can be updated by $\lambda^{(l)} = \frac{\sum_{m=1}^{M}\sum_{k=1}^{K_m} R_{m,k}^{sec\ (l)}}{\sum_{m=1}^{M}\sum_{k=1}^{K_m} P_{m,k}^{(l)} + \sum_{m=1}^{M}\sum_{k=0}^{K_m} Q_{m,k}^{(l)} + P_f}$, where $R_{m,k}^{sec\ (l)}$, $P_{m,k}^{(l)}$ and $Q_{m,k}^{(l)}$ are the updated rates and power after solving (5.36). As shown in [42], $\lambda^{(l)}$ keeps growing as l increases. When $\lambda^{(l)} - \lambda^{(l-1)}$ is smaller than a certain threshold, e.g., 10^{-3}, the iterations terminate, and the obtained $\lambda^{(l)}$ is the maximum EE of (5.24).

Then, the problem lies in how to solve (5.36) for a given λ. It is clear that (5.36) is similar to the sum rate maximization problem (5.22), except for the extra linear part in the objective function. Adding a linear part does not affect the way of solving the problem, and thus, we can apply the proposed sum rate maximization here directly. The specific procedure is summarized in Algorithm 6.

Algorithm 6 Energy-Efficient Power Allocation Algorithm.

Initialize $l \leftarrow 0, \varepsilon^{\star} \leftarrow 1, \varepsilon \leftarrow 10^{-3}$; $\lambda \leftarrow 0$; Initialize feasible power $\mathbf{P}^{(0)}$;
while $\varepsilon^{\star} \geq \varepsilon$
  while $|\mathbf{P}^{(l)} - \mathbf{P}^{(l-1)}| > 10^{-3}$ or $|\mathbf{Q}^{(l)} - \mathbf{Q}^{(l-1)}| > 10^{-3}$
    $\mathbf{P}^{(l)}, \mathbf{Q}^{(l)} \leftarrow \max \sum_{m=1}^{M} \sum_{k=1}^{K_m} R_{m,k}^{sec} - \lambda(\sum_{m=1}^{M} \sum_{k=1}^{K_m} P_{m,k} + \sum_{m=1}^{M} \sum_{k=0}^{K_m} Q_{m,k} + P_f)$ s.t. (5.22b), (5.22c), (5.22d)
  end while
  $\varepsilon^{\star} \leftarrow \sum_{m=1}^{M} \sum_{k=1}^{K_m} R_{m,k}^{sec\ (l)} - \lambda(\sum_{m=1}^{M} \sum_{k=1}^{K_m} P_{m,k}^{(l)} + \sum_{m=1}^{M} \sum_{k=0}^{K_m} Q_{m,k}^{(l)} + P_f)$;
  $\lambda \leftarrow \frac{\sum_{m=1}^{M} \sum_{k=1}^{K_m} R_{m,k}^{sec\ (l)}}{\sum_{m=1}^{M} \sum_{k=1}^{K_m} P_{m,k}^{(l)} + \sum_{m=1}^{M} \sum_{k=0}^{K_m} Q_{m,k}^{(l)} + P_f}$;
  $l \leftarrow l + 1$;
end while
if $R_{m,k}^{sec\ (l)} \leftarrow 0, \forall m \in \{1, \cdots, M\}, k \in \{1, \cdots, K_m\}$
  $R_{m,k}^{sec\ (l)} \leftarrow 0$;
end
$\eta_{\mathrm{EE}} \leftarrow \frac{\sum_{m=1}^{M} \sum_{k=1}^{K_m} R_{m,k}^{sec\ (l)}}{\sum_{m=1}^{M} \sum_{k=1}^{K_m} P_{m,k}^{(l)} + \sum_{m=1}^{M} \sum_{k=0}^{K_m} Q_{m,k}^{(l)} + P_f}$.

5.6.3 Complexity and Convergence

The proposed SE maximization algorithm includes inner and outer iterations. For the inner iteration, i.e., the DC programming, its convergence has been shown in [39, 40]. For the outer iteration, on one hand, the SE increases or remains unchanged for both the uplink and downlink PA; on the other hand, there exists an upper bound for the SE. Therefore, the outer iteration terminates within a limited number of iterations, i.e., the proposed SE maximization algorithm always converges.

The proposed EE maximization algorithm also includes inner and outer iterations. For the inner iteration, i.e., the SE maximization, its convergence has been shown above. For the outer iteration, i.e., the fractional programming, it always converges to the stationary and optimal solution [42]. Therefore, the proposed EE maximization algorithm always converges.

Now, we discuss the computational complexity of the proposed algorithms. First, we look at the proposed SE maximization algorithm. Denote the number of iterations for

solving the uplink and downlink PA as I_1 and I_2, respectively. The corresponding number of dual variables for solving (5.30) and (5.35) is denoted as D_1 and D_2, respectively. Then, if the number of outer iteration is I_3, the overall computational complexity of the proposed SE maximization algorithm is $O\left(I_3(I_1D_1^2 + I_2D_2^2)\right)$. Next, we consider the proposed EE maximization problem. Denote its outer iteration as I_4, then it can be easily shown that the overall computational complexity of the proposed EE maximization algorithm is $O\left(I_4I_3(I_1D_1^2 + I_2D_2^2)\right)$.

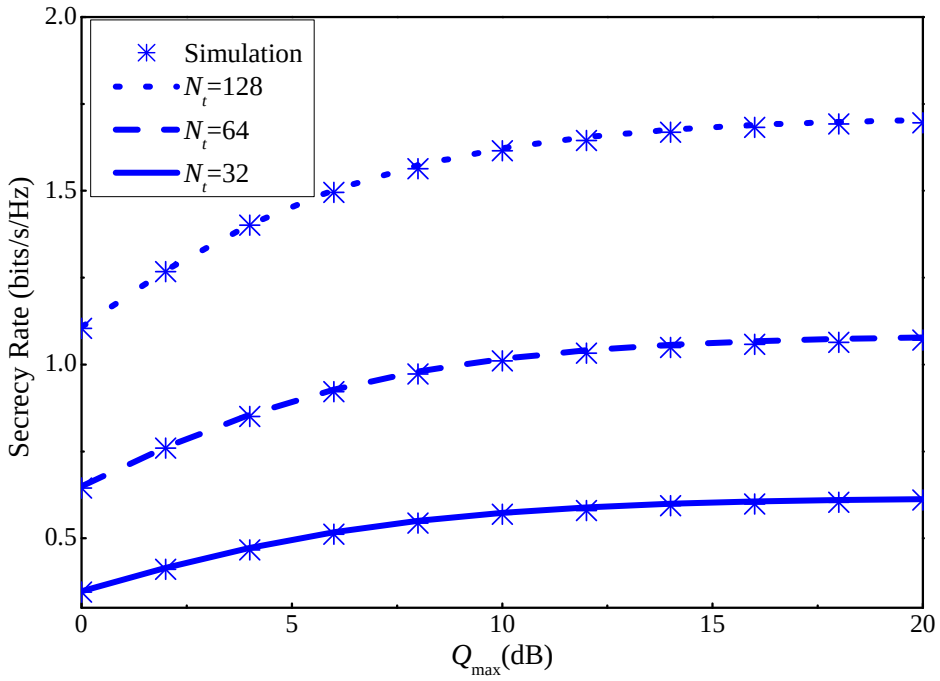


Fig. 5.2: Secrecy rate at the 2^{nd} user in the 5^{th} cluster versus the total transmit power at the BS, for different numbers of transmit antennas.

5.7 Numerical Results

In this section, the behavior of the system without PA is we first investigated to highlight the effects of key parameters on the secrecy performance in subsection 5.7.1. The effectiveness of our proposed PA algorithms is then evaluated in subsection 5.7.2.

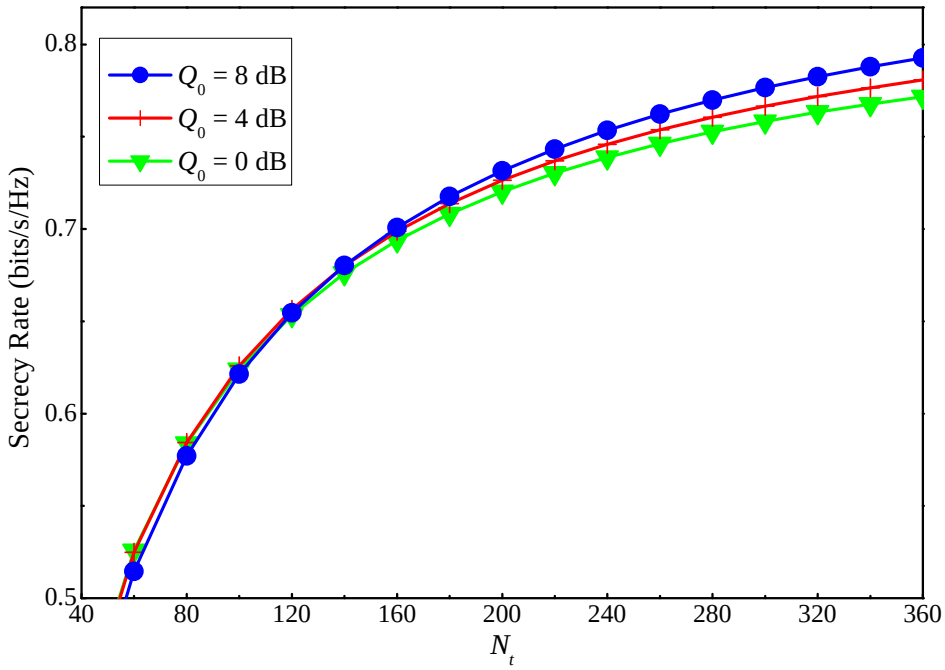


Fig. 5.3: Secrecy rate at the 2^{nd} user in the 5^{th} cluster versus the number of antennas at the BS, for different AN powers.

5.7.1 Fixed PA

Without loss of generality, the following scenario is considered. The total transmit power is allocated 80% for information transmission and 20% for AN. The effect of varying the AN power allocation will be shown later in Fig. 5.3. The power is equally assigned to each user, and the AN power for each cluster is the same. $T = 300$ units and $\tau = M$ units. $\beta_{m,k}$ for each user is a random value between 0 and 100 and satisfies the condition $\beta_{m,1} \geq ... \geq \beta_{m,K_m}$, while that for the illegitimate user is fixed to β_E =10. Unless explicitly mentioned, this setup is kept throughout the section.

Without loss of generality, the ergodic secrecy rate of the 2^{nd} user in the 5^{th} cluster is selected to show in Fig. 5.2. The number of cluster is $M = 10$ and the number of users in a cluster is $K = 2$.[4] It can be seen that the approximation in (5.11) and the simulation results match very well. Throughout the numerical results section, this approximation

[4]The subscript m in K_m is dropped since the same number of users is considered in clusters.

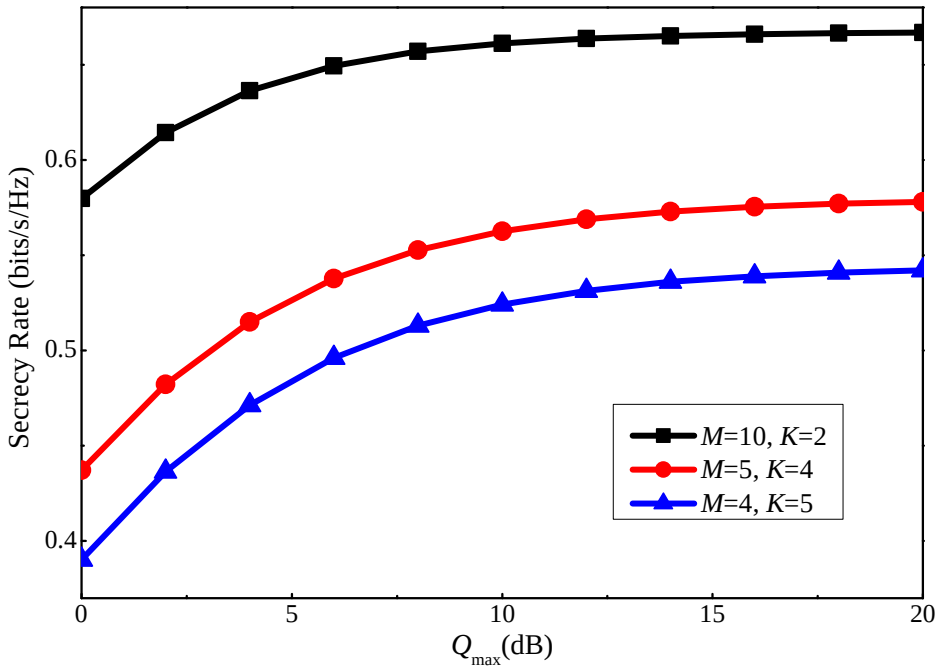


Fig. 5.4: Secrecy rate at the 2^{nd} user of the 2^{nd} cluster versus the total transmit power at the BS, for different clustering scenarios.

will be used. Besides, when the total transmit power at the BS increases, the secrecy rate at a user converges to a constant value. This is because of the interference and AN within the cluster and from other clusters. In addition, we can also observe that an increase in the number of antennas at the BS can lift the secrecy performance. The reason is that by increasing the number of antennas, the spatial transmitting beams become sharper, which leads to a decrease in inter-cluster interference and AN leakage, and an increase in the desired signal. The next figure will reveal how to take advantage of this property to enhance secrecy performance.

Figure 5.3 demonstrates the advantage of combining AN and massive MIMO technique in NOMA networks. In this setup, the transmit power assigned to each user is 10 dB, and the AN power is varied as $\{0, 4, 8\}$ dB. The number of clusters is $M = 10$ with $K = 2$ users in a cluster. We can observe that when the number of transmit antennas is sufficiently large, the more the AN allocated power is, the better the secrecy performance at the 2^{nd}

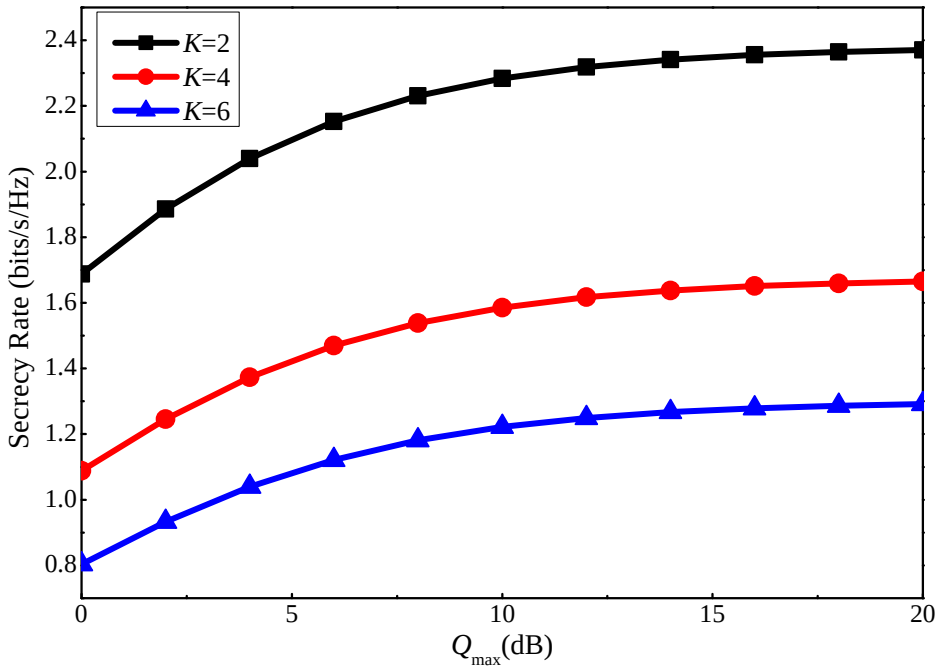


Fig. 5.5: The secrecy rate at the the 2^{nd} cluster versus the total transmit power at the BS, for a fixed number of clusters and different numbers of users.

user is. The main reason is that for the legitimate side, the channel hardening property of massive MIMO technique helps reducing the AN leakage and the inter-cluster interference at each cluster. Meanwhile, the secrecy performance of the eavesdropper decreases when the AN power increases.

Figures 5.4 and 5.5 depict the effect of clustering on the secrecy performance. In Fig. 5.4, the total number of users is 20, which are clustered into three scenarios: $\{M = 10, K = 2\}$, $\{M = 5, K = 4\}$, and $\{M = 4, K = 5\}$. The total transmit power for each scenario is the same. The results show that the smaller the number of users in a cluster is, the better the secrecy performance at a user is. Meanwhile, in Fig. 5.5, the scenario of limited number of orthogonal sequences is shown. In this scenario, we assume that the number of available orthogonal sequences is 10, therefore, the number of clusters is $M = 10$. The number of users in a cluster is varied as $K = \{2, 4, 6\}$ to highlight its effect on the secrecy performance of a cluster. It is observed that although the transmit power

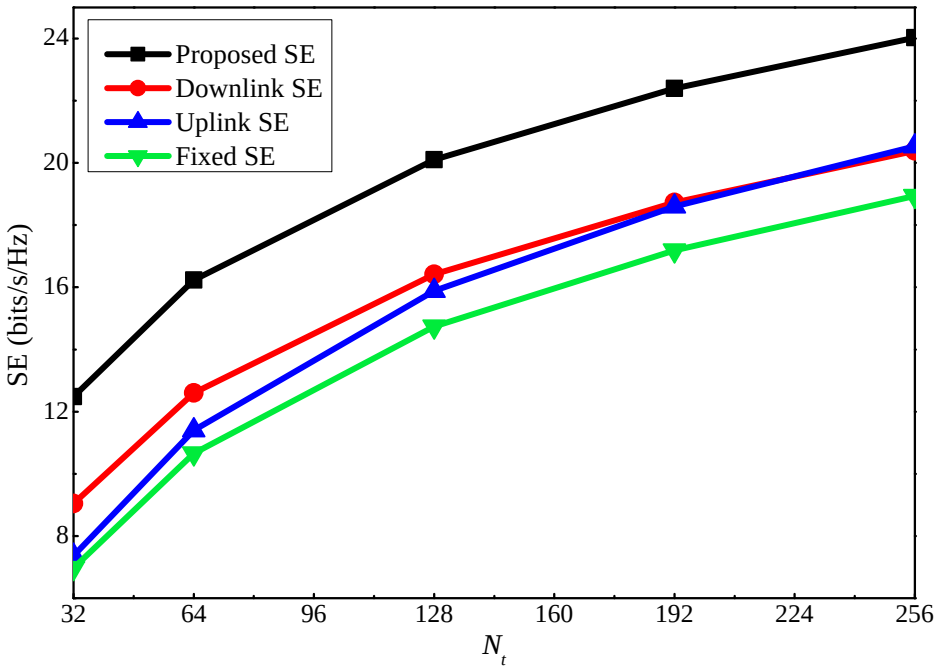


Fig. 5.6: SE comparison between the proposed algorithm and baseline algorithms.

for each user is identical and the AN power for each cluster is the same, the cluster with more users has smaller total secrecy rate than the ones with a smaller number of users. The reason is that when the number of users in a cluster is small, the error of the uplink training process at this cluster is also small. As a consequence, the beam of the BS for this cluster is more precise, followed by a decrease in intra-cluster interference and AN leakage. This also reduces the imposed interference from this cluster to other clusters. In other words, for a better secrecy performance of each user and cluster, it is crucial to keep the number of users in a cluster small (minimum is two users for NOMA networks).

5.7.2 Optimized PA

In the following, the effectiveness of the proposed SE and EE maximization algorithms is investigated. A scenario with four clusters, each with three users, i.e., $M = 4$ and $K = 3$ is considered. The simulation parameters are as follows: $Q_{\max} = 20$ dB, $P_{\max} = 0$ dB.

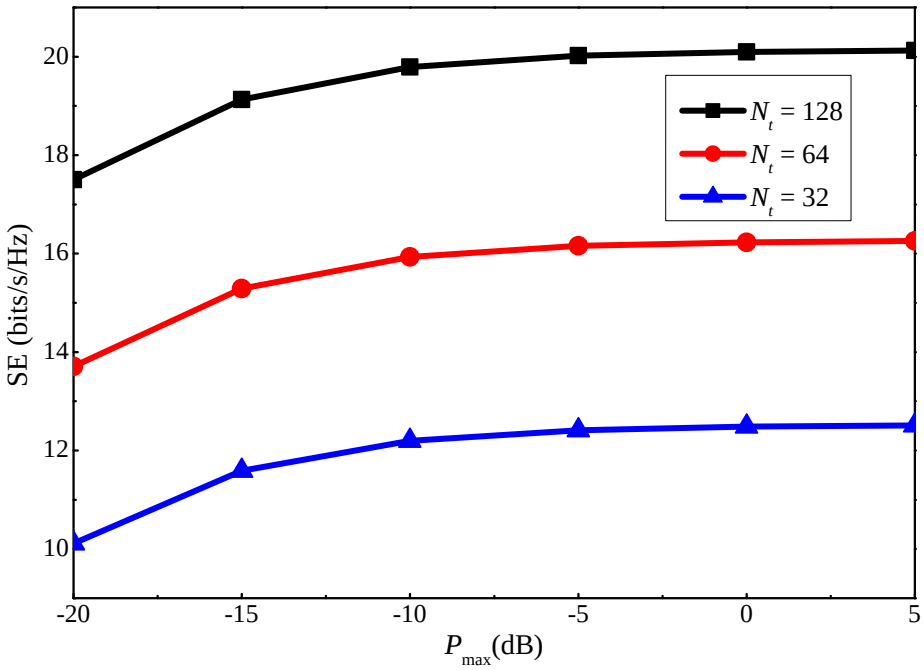


Fig. 5.7: SE versus the maximum uplink power, for different numbers of BS antennas.

$T = 300$ units. The large scale channel gain $\beta_{m,k}$ for each user is a random value between 0 and 100, while that for the illegitimate user is fixed to β_E =10.

First, the effectiveness of the proposed SE maximization algorithm, referred to as Proposed SE is investigated. It is compared with three baseline algorithms, as follows: Downlink SE, which allocates the maximum uplink power to each user, and on this basis, performs PA for the downlink transmission as the Proposed SE. In contrast, the Uplink SE first allocates 80% of the total downlink power to the users equally, and 20% of the total downlink power to the AN equally. Then, uplink power is optimized as the Proposed SE. Fixed SE allocates the maximum uplink power for each user, equal downlink power allocation among the users, and the AN as in the above subsection. As shown in Fig. 5.6, the SE provided by all four algorithms grows with the number of transmit antennas. Moreover, among them, it can be seen that Proposed SE achieves the best performance, followed by Downlink SE, Uplink SE, and Fixed PA. This fully reveals the necessity of performing power optimization for the considered system. Furthermore, both uplink and

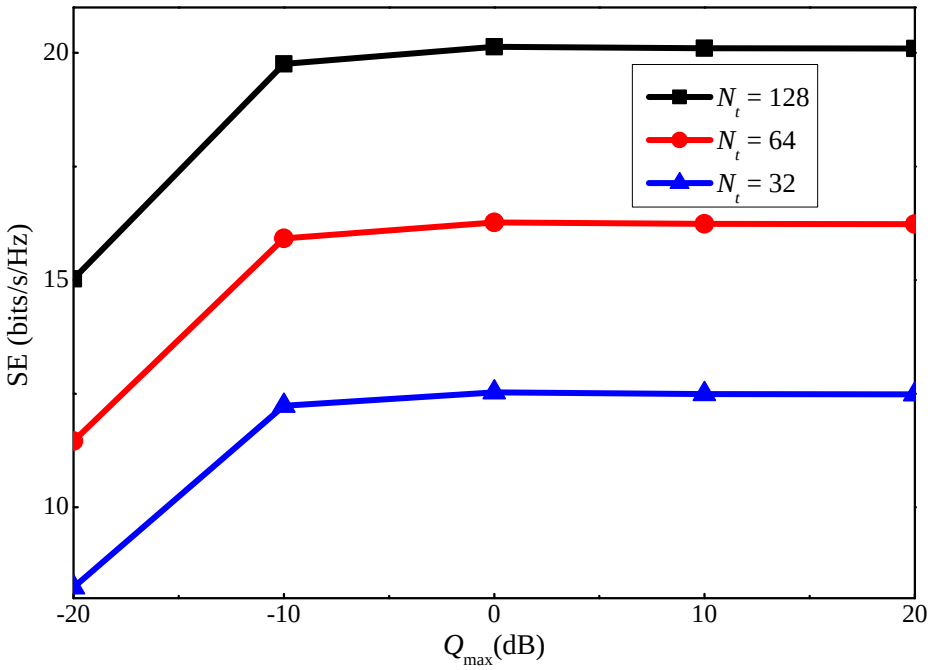


Fig. 5.8: SE versus the maximum downlink power, for different numbers of BS antennas.

downlink PA are required to achieve the best performance. Nonetheless, by comparing Downlink SE and Uplink SE, we can conclude that an appropriate allocation of the downlink power may play a larger role in the current setting.

To further show the effect of the uplink and downlink power on the achieved SE, Figs. 5.7 and 5.8 plot the SE versus the maximum uplink and downlink power, respectively. $N_t = \{32, 64, 128\}$ is respectively considered in each case. It is clear that the SE increases with both the maximum uplink and downlink powers. The former is because increasing the maximum uplink power leads to a more precise channel estimation result, which improves the beamforming sharpness and thus, the SE. The latter is because more power is available for data transmission. However, after a certain point, the increase becomes minor for both power values. This can be explained by the logarithmic relation between the power and user rate. Moreover, for the downlink power, increasing it also leads to a larger illegitimate rate and intra-cluster interference. Besides, by comparing the three antenna scenarios, we can conclude that increasing the number of antennas can

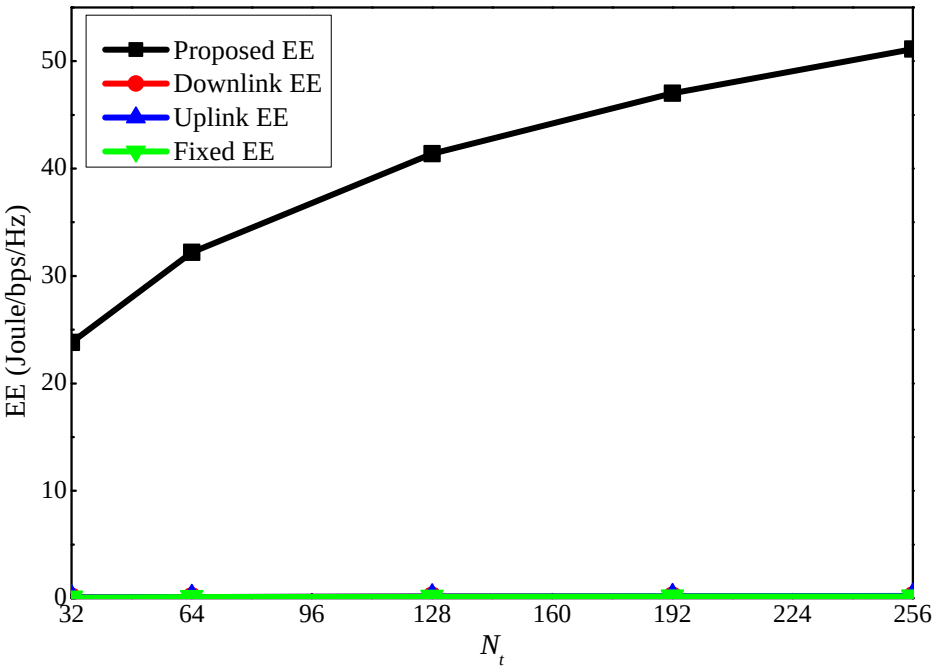


Fig. 5.9: EE comparison between the proposed algorithm and other baseline algorithms.

significantly increase the SE.

Next, the proposed EE algorithm is investigated. Here $P_f = -5$ dB. The proposed EE maximization algorithm is first compared with the other three baseline algorithms when $Q_{\max} = 20$ dB and $P_{\max} = 0$ dB. According to Fig. 5.9, the EE for the other algorithms is quite small compared with the proposed algorithm. This is because when $Q_{\max} = 20$ dB and $P_{\max} = 0$ dB, the power level is quite high, and thus, a large part of the available power is not used to maximize the EE. However, for the three baseline algorithms, at least one of the uplink and downlink power is fully consumed according to the setting. This leads to low EE. Figure 5.10 only shows these three algorithms, and it can be seen that all of them increase with the antenna number as the proposed algorithm.

Similar to the SE, how the EE varies with the maximum uplink and downlink power is shown in Figs. 5.11 and 5.12, respectively. For both cases, the EE first grows with the maximum power constraint, and after a certain threshold, i.e., $P_{\max} = -20$ dB and $Q_{\max} = -10$ dB, it remains unchanged even if the maximum power constraint continues

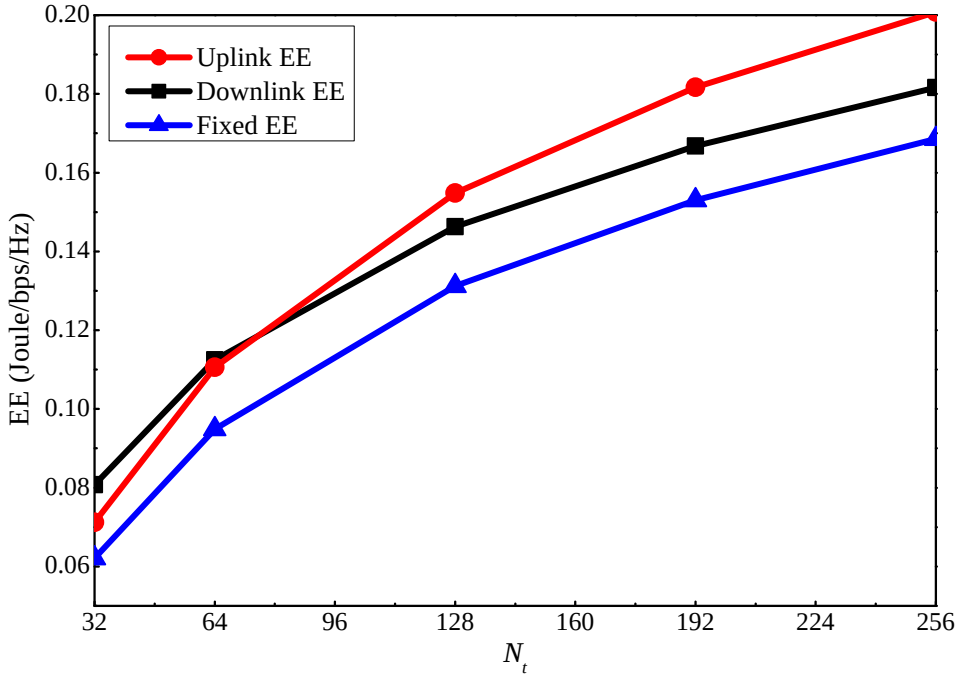


Fig. 5.10: EE for the three baseline algorithms.

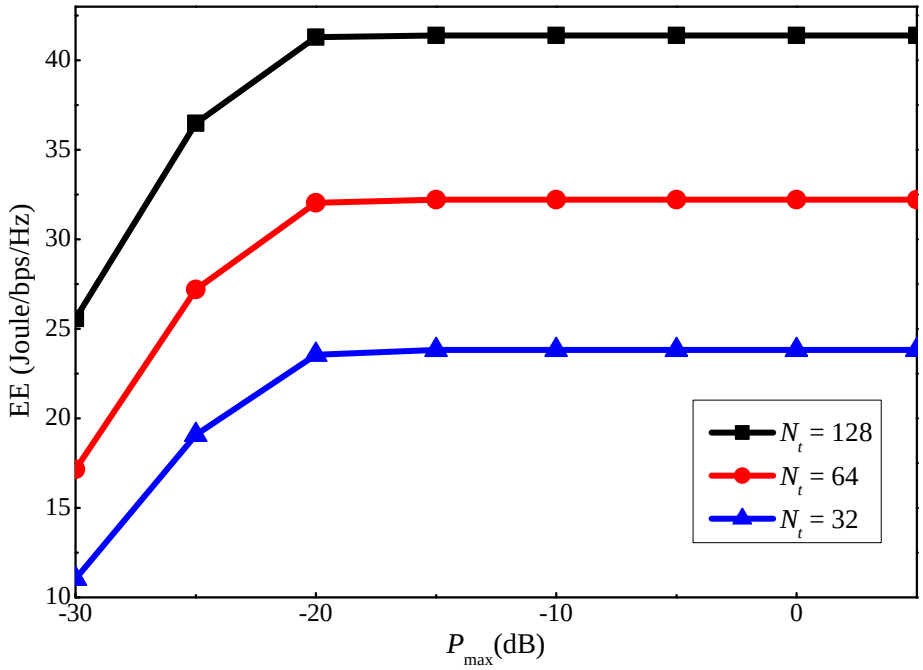


Fig. 5.11: EE versus the maximum uplink power, for different numbers of BS antennas.

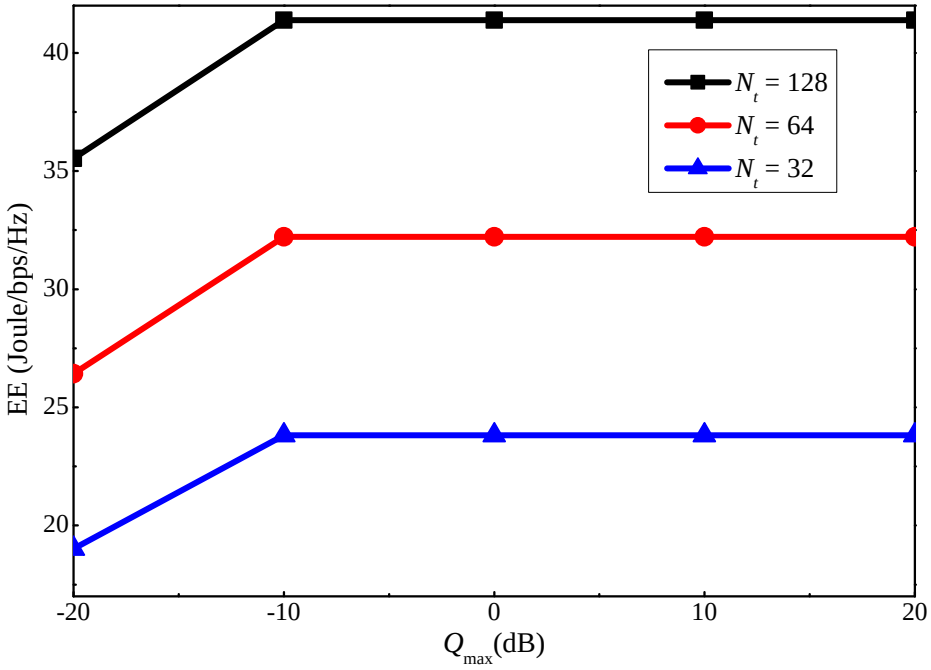
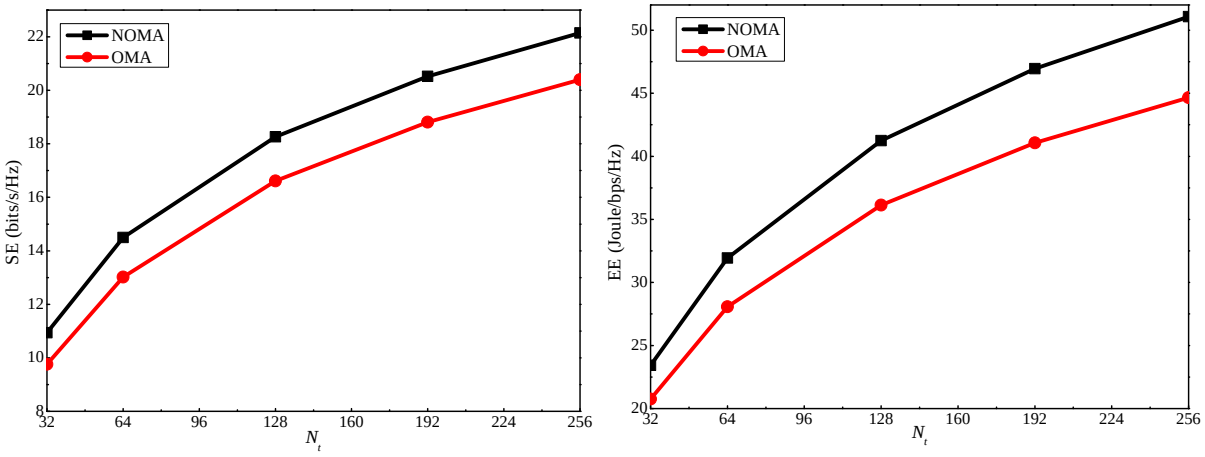


Fig. 5.12: EE versus the maximum downlink power, for different numbers of BS antennas.



(a) (b)

Fig. 5.13: Performance comparison for NOMA and OMA when the number of antenna varies: (a) SE; (b) EE.

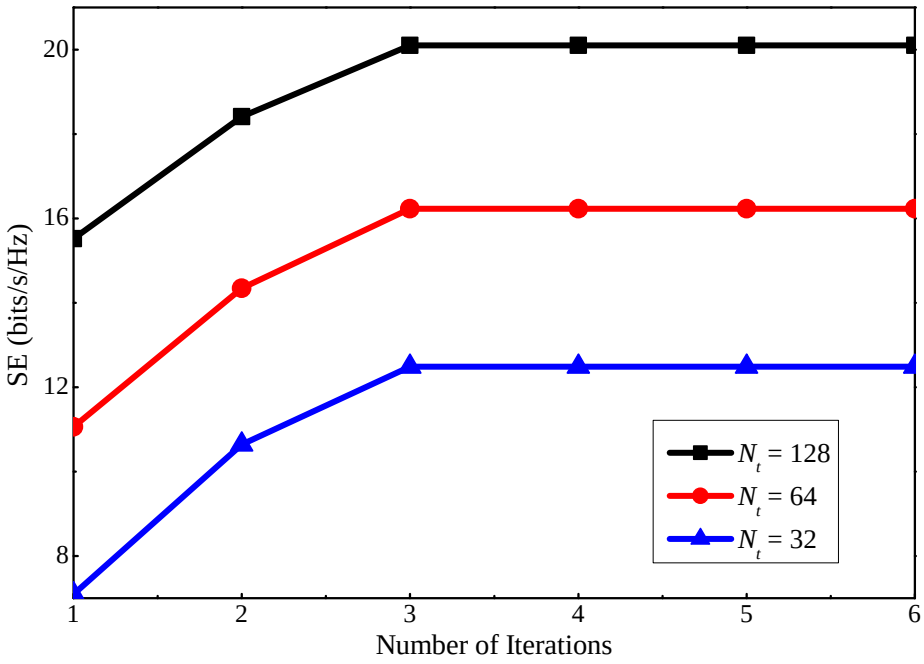


Fig. 5.14: Convergence of the proposed SE algorithm.

to grow. This is because the slow increases in the SE cannot compensate for the power increment when the power is high, and thus, no more power will be consumed by the users to maximize the EE. By comparing the EE figures with the sum rate ones, i.e., Fig. 5.7 versus Fig. 5.11, and Fig. 5.8 versus Fig. 5.12, we can observe that the EE reaches the turning point at a smaller power value than the sum rate. This is because after the sum rate increment over the power declines to a certain value, no more extra power is used to maximize the EE.

The baseline massive MIMO-OMA can be considered as a special case of the proposed massive MIMO-NOMA scheme with just one user in each cluster. Accordingly, the legitimate achievable rate of the m-th user is:

$$R_m^{OMA} = \left(1 - \frac{\tau}{T}\right) \log_2 \left(1 + \frac{\kappa_m}{\sum_{i=1}^{3} I_{m,i} + 1}\right), \tag{5.37}$$

where $\kappa_m = Q_m\beta_m\rho_m N_t$, $I_{m,1} = Q_m\beta_m(1-\rho_m)$, $I_{m,2} = \sum_{i\neq m}^{M} Q_i\beta_m$, $I_{m,3} = Q_{m,0}\beta_m(1-\rho_m) + \sum_{i\neq m}^{M} Q_{i,0}\beta_m$, Q_m is the downlink power for the m-th user, $Q_{m,0}$ is the AN power for the m-th user, β_m is the large scale fading of the m-th user, $\rho_m = \frac{P_m\beta_m\tau}{P_m\beta_m\tau+1}$, τ is the length of training sequences that is the same as the NOMA case, and P_m is the uplink transmit power of the m-th user. The achievable eavesdropping rate corresponding to the m-th user is:

$$R_{E,m}^{OMA} = \left(1-\frac{\tau}{T}\right)\log_2\left(1+\frac{Q_m\beta_E}{\sum_{i\neq m}^{M} Q_i\beta_E + \sum_{i=1}^{M} Q_{i,0}\beta_E + 1}\right). \tag{5.38}$$

The achievable secrecy rate of the m-th user is

$$R_{S,m}^{OMA} = [R_m^{OMA} - R_{E,m}^{OMA}]^+. \tag{5.39}$$

In simulations, to compare the proposed massive MIMO-NOMA with the baseline massive MIMO-OMA, a scenario with four clusters and two users in each cluster is considered. TDMA is used for the baseline massive MIMO-OMA, and thus, each user in one cluster is only served half the time. Fig. 5.13 shows the corresponding SE and EE comparison between the considered schemes. It is clear that the proposed scheme outperforms the baseline massive MIMO-OMA when the number of antennas at the BS increases, which shows its superiority.

Finally, Figs. 5.14 and 5.15 show how many iterations are required for the proposed SE and EE maximization algorithms to converge, respectively. Note that here an iteration means solving either the uplink or the downlink DC programming problem, which requires to solve an average of five convex problems according to the simulation. Results for three different antenna numbers are presented when $Q_{\max} = 20$ dB and $P_{\max} = 0$ dB. It can be seen that a small number of iterations are required for the proposed SE and EE maximization algorithms to converge.

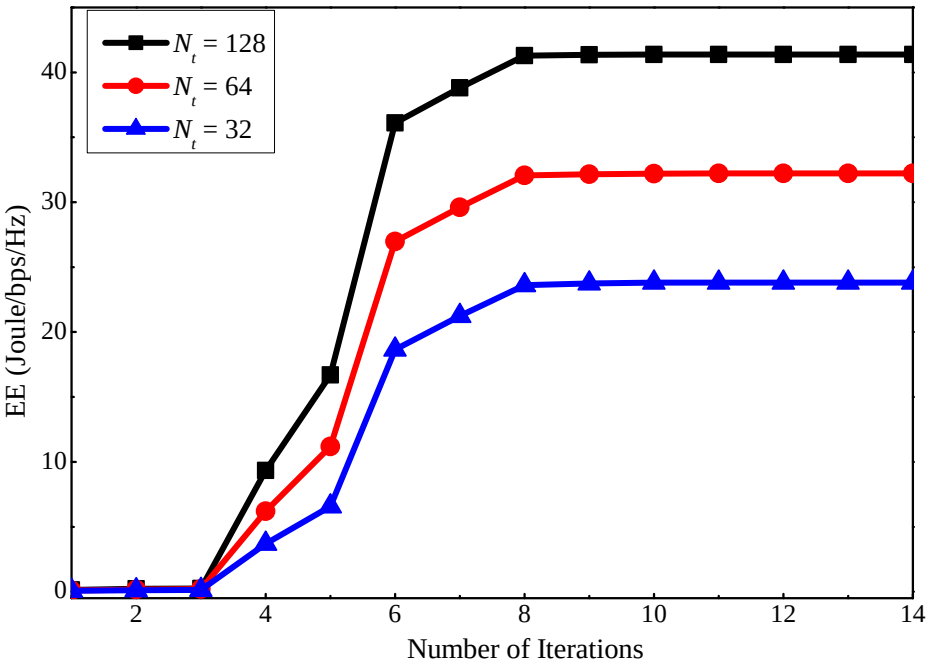


Fig. 5.15: Convergence of the proposed EE algorithm.

5.8 Conclusion

In this chapter, an AN-aided scheme has been proposed to ensure secrecy in massive MIMO-NOMA networks. The ergodic secrecy rate and its asymptotic value have been derived to spotlight the roles of key parameters on the secrecy performance of the considered system. The results have revealed that with a sufficiently large number of transmit antennas at the BS, only the illegitimate side is affected by the AN. In addition, when the transmit power at the BS is high, the secrecy performance of a user is independent of the inter-cluster interference and AN and is determined by the uplink training process, which depends on the number of users in a cluster, the uplink transmit power, and the large-scale fading. Besides, the results also suggest to keep the number of users in a cluster small for a better secrecy performance at each user and cluster. Furthermore, numerical results validate that our proposed optimization algorithms can obtain significant improvements over the baseline algorithms, i.e., Uplink PA, Downlink PA and Fixed PA, in terms of

the sum ergodic secrecy rate and energy efficiency. This fully reveals the necessity of performing power optimization for the considered system, and the effectiveness of the proposed algorithms. Finally, from the perspective of sum ergodic secrecy rate and its energy efficiency, the proposed system surpasses the conventional massive MIMO-OMA system.

References

[1] V. W. S. Wong et al., *Key Technologies for 5G Wireless Systems.* Cambridge, UK: Cambridge University Press, 2017.

[2] S. M. R. Islam, M. Zeng, and O. A. Dobre, "NOMA in 5G systems: Exciting possibilities for enhancing spectral efficiency," *IEEE 5G Tech. Focus,* vol. 1, no. 2, May 2017. [Online]. Available: 307 http://5g.ieee.org/tech-focus.

[3] S. M. R. Islam, M. Zeng, O. A. Dobre, and K. Kwak, "Resource allocation for downlink noma systems: Key techniques and open issues," *IEEE Wireless Commun. Mag.*, vol. 25, no. 2, pp. 40–47, April 2018.

[4] S. M. R. Islam et al., "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surv. Tuts.*, vol. 19, no. 2, pp. 721–742, Second quarter 2017.

[5] M. Zeng, G. I. Tsiropoulos, O. A. Dobre, and M. H. Ahmed, "Power allocation for cognitive radio networks employing non-orthogonal multiple access," in *Proc. IEEE Globecom*, Washington DC, USA, Dec. 2016.

[6] Z. Wei, D. W. K. Ng, J. Yuan, and H. Wang, "Optimal resource allocation for power-efficient mc-noma with imperfect channel state information," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3944–3961, Sep. 2017.

[7] 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, Oct. 2017.

[8] ——, "On the sum rate of MIMO-NOMA and MIMO-OMA systems," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 534–537, Aug. 2017.

[9] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[10] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "A fair individual rate comparison between MIMO-NOMA and MIMO-OMA," in *Proc. IEEE Globecom Wkshps*, Singapore, Dec 2017, pp. 1–5.

[11] N. Yang, L. Wang, G. Geraci, M. Elkashlan, J. Yuan, and M. D. Renzo, "Safeguarding 5G wireless communication networks using physical layer security," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 20–27, Apr. 2015.

[12] A. D. Wyner, "The wire-tap channel," *Bell Syst. Tech. J.*, vol. 54, no. 8, pp. 1355–1387, Jan. 1975.

[13] Y.-S. Shiu, S. Chang, H.-C. Wu, S. Huang, and H.-H. Chen, "Physical layer security in wireless networks: A tutorial," *IEEE Wireless Commun.*, vol. 18, no. 2, pp. 66–74, Apr. 2011.

[14] J. Chen, L. Yang, and M.-S. Alouini, "Physical layer security for cooperative NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4645–4649, May 2018.

[15] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and L. Hanzo, "Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1656 – 1672, Mar. 2017.

[16] Y. Zhang, H.-M. Wang, Q. Yang, and Z. Ding, "Secrecy sum rate maximization in non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 930–933, May 2016.

[17] B. He, A. Liu, N. Yang, and V. K. N. Lau, "On the design of secure non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2196–2206, Oct. 2017.

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

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

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

[21] W. Hao, M. Zeng, Z. Chu, S. Yang, and G. Sun, "Energy-efficient resource allocation for mmWave massive MIMO HetNets with wireless backhaul," *IEEE Access*, vol. 6, pp. 2457–2471, Feb. 2018.

[22] W. Hao, M. Zeng, Z. Chu, and S. Yang, "Energy-efficient power allocation in millimeter wave massive MIMO with non-orthogonal multiple access," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 782–785, Dec. 2017.

[23] M. Zeng, W. Hao, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation in uplink mmWave massive MIMO with NOMA," *IEEE Trans. Veh. Technol.*, pp. 1–1, 2019.

[24] J. Ma, C. Liang, C. Xu, and L. Ping, "On orthogonal and superimposed pilot schemes in massive MIMO NOMA systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2696 – 2707, Dec. 2017.

[25] X. Chen, Z. Zhang, C. Zhong, D. W. K. Ng, and R. Jia, "Exploiting inter-user interference for secure massive non-orthogonal multiple access," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 4, pp. 788–801, Apr. 2018.

[26] J. Zhu, R. Schober, and V. K. Bhargava, "Linear precoding of data and artificial noise in secure massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 2245–2261, Mar. 2016.

[27] N.-P. Nguyen, H. Q. Ngo, T. Q. Duong, H. D. Tuan, and K. Tourki, "Secure massive MIMO with the artificial noise-aided downlink training," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 4, pp. 802 – 816, Apr. 2018.

[28] Y.-Y. Zhang, J.-K. Zhang, and H.-Y. Yu, "Physically securing energy-based massive MIMO MAC via joint alignment of multi-user constellations and artificial noise," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 4, pp. 829 – 844, Apr. 2018.

[29] N.-P. Nguyen, H. Q. Ngo, T. Q. Duong, H. D. Tuan, and D. B. da Costa, "Full-duplex cyber-weapon with massive arrays," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5544 – 5558, Aug. 2017.

[30] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. San Diego, CA: Academic press, 2007.

[31] H. Tabassum, E. Hossain, and J. Hossain, "Modeling and analysis of uplink non-orthogonal multiple access in large-scale cellular networks using poisson cluster processes," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3555–3570, Aug 2017.

[32] H. Sun, B. Xie, R. Q. Hu, and G. Wu, "Non-orthogonal multiple access with sic error propagation in downlink wireless mimo networks," in *Proc. IEEE VTC*, Sep. 2016, pp. 1–5.

[33] X. Chen, Z. Zhang, C. Zhong, R. Jia, and D. W. K. Ng, "Fully non-orthogonal communication for massive access," *IEEE Trans. Commun.*, vol. 66, no. 4, pp. 1717–1731, Apr. 2018.

[34] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation for MIMO-NOMA with multiple users in a cluster," *IEEE Access*, vol. 6, pp. 5170–5181, 2018.

[35] ——, "Energy-efficient power allocation for hybrid multiple access systems," in *Proc. IEEE ICC Wkshps*, Kansas City, MO, USA, May 2018, pp. 1–5.

[36] W. Hao, Z. Chu, F. Zhou, S. Yang, G. Sun, and K. Wong, "Green communication for NOMA-based CRAN," *IEEE Internet of Things J.*, pp. 1–1, 2018.

[37] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation for uplink NOMA," in *Proc. IEEE Globecom*, Abu Dhabi, UAE, Dec. 2018, pp. 1–6.

[38] ——, "Energy-efficient joint User-RB association and power allocation for uplink hybrid NOMA-OMA," *IEEE Internet of Things J.*, pp. 1–1, 2019.

[39] H. H. Kha, H. D. Tuan, and H. H. Nguyen, "Fast global optimal power allocation in wireless networks by local d.c. programming," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 510–515, Feb. 2012.

[40] N. Vucic, S. Shi, and M. Schubert, "DC programming approach for resource allocation in wireless networks," in *Proc. Int. Symp. Modeling Optimization Mobile, Ad Hoc Wireless Netw.*, May 2010, pp. 380–386.

[41] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21." *Available: http://cvxr.com/cvx*, Dec. 2010.

[42] W. Dinkelbach, "On nonlinear fractional programming," *Manag. Sci.*, vol. 13, no. 7, pp. 3492–498, Mar. 1967.

Chapter 6

Conclusions

In this final chapter, the contributions presented in the dissertation are summarized and several potential extensions of the work are discussed.

6.1 Conclusions

The following conclusions can be drawn from the dissertation:

- The capacity of MIMO-NOMA was compared with that of MIMO-OMA, when multiple users are grouped into a cluster. The superiority of MIMO-NOMA over MIMO-OMA was demonstrated in terms of both sum channel capacity and ergodic sum capacity. It was also proved that the more users are admitted to the same cluster, the lower is the achieved sum rate, which implies a tradeoff between sum rate and number of admitted users. On this basis, a user admission scheme was proposed, which achieves optimal results in terms of both sum rate and number of admitted users when the SINR thresholds of the users are equal. When the SINR thresholds of the users are different, the proposed scheme still achieves good performance in balancing both criteria. Furthermore, the proposed scheme is of low

complexity, i.e., linear in the number of users in each cluster.

- The EE maximization problem was studied for a multi-cluster multi-user MIMO-NOMA system under a QoS constraint for each user. An optimal PA strategy was proposed to solve the considered EE maximization problem. Numerical results showed that the proposed PA strategies outperform OMA and equal power NOMA in terms of EE, which verified their effectiveness.

- The energy-efficient resource allocation was addressed for HMA uplink with QoS requirements for each user. Based on swap matching in many-to-one bipartite graph, a joint user-RB association and power allocation scheme was proposed, which is guaranteed to converge. Under a given user-RB association, it was shown that the system EE maximization equals cluster EE maximization. Then, the feasibility conditions were derived, and the EE maximization was solved using Dinkelbach's algorithm. Moreover, to further relieve the computational burden, a low-complexity optimal algorithm was proposed for solving the convex optimization subproblem inside the Dinkelbach's algorithm. For the two user case, analytical solutions were derived for the two SIC orders.

- An AN-aided scheme was proposed to ensure secrecy in massive MIMO-NOMA networks. The ergodic secrecy rate and its asymptotic value were derived to spotlight the roles of key parameters on the secrecy performance of the considered system. The results revealed that with a sufficiently large number of transmit antennas at the BS, only the illegitimate side is affected by the AN. In addition, when the transmit power at the BS is high, the secrecy performance of a user is independent of the inter-cluster interference and AN and is determined by the uplink training process, which depends on the number of users in a cluster, the uplink transmit power, and the large-scale fading. Besides, the results also suggest to keep the number of

users in a cluster small for a better secrecy performance at each user and cluster. Furthermore, numerical results validated that the proposed optimization algorithms can obtain significant improvements over the baseline algorithms, i.e., Uplink PA, Downlink PA and Fixed PA, in terms of the sum ergodic secrecy rate and energy efficiency.

6.2 Possible Directions of Research

As the work on MIMO-NOMA is only at an early stage, there can be different directions to extend our work, which can be briefly outlined as follows:

- The works devoted to MIMO-NOMA consider the PA under given user clustering; it may be worth investigating the joint user clustering and PA allocation as for the uplink HMA.

- The work on massive MIMO-NOMA performs channel estimation assuming that user clustering is already done. However, user clustering cannot be performed properly without knowing CSI, i.e., channel estimation. As a result, channel estimation and user clustering seem to be a chicken-and-egg problem, which is non-trivial to handle. Studying such a problem is of practical interest.

- Most works on NOMA assume perfect CSI at the transmitter and SIC at the receiver. It is worthy to investigate the resource allocation and performance evaluation of NOMA under more practical scenarios, such as statistical CSI at the transmitter and imperfect SIC at the receiver. This is expected to facilitate the actual deployment of NOMA in 5G and beyond networks.

- The integration of NOMA into other advanced wireless technologies seems promis-

ing and is worth of investigation. Incipient studies have shown that NOMA can contribute to wireless caching, grant-free access for Internet-of-Things, unmanned aerial vehicle communication, mmWave communication and so on. There are plenty of opportunities to study the role of NOMA in these research areas.

References

Chapter 1

[1] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication.* Cambridge, UK: Cambridge University Press, 2005.

[2] L. Dai et al., "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[3] S. M. R. Islam et al., "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surv. Tuts.*, vol. 19, no. 2, pp. 721–742, Second quarter 2017.

[4] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[5] S. M. R. Islam, M. Zeng, and O. A. Dobre, "NOMA in 5G systems: Exciting possibilities for enhancing spectral efficiency," *IEEE 5G Tech. Focus,* vol. 1, no. 2, May 2017. [Online]. Available: 307 http://5g.ieee.org/tech-focus.

[6] S. M. R. Islam, M. Zeng, O. A. Dobre, and K. Kwak, "Resource allocation for downlink noma systems: Key techniques and open issues," *IEEE Wireless Commun. Mag.*, vol. 25, no. 2, pp. 40–47, April 2018.

[7] Q. Sun, S. Han, I. Chin-Lin, and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 4, pp. 405–408, Dec. 2015.

[8] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, Jun. 2016.

[9] 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, Aug. 2017.

[10] ——, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Select. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.

[11] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[12] Y. Saito et al., "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE Veh. Technol. Conf.*, Dresden, Germany, Jun. 2013, pp. 1–5.

[13] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[14] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Power and channel allocation for non-orthogonal multiple access in 5G systems: Tractability and computation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8580–8594, Dec. 2016.

[15] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686–7698, Nov. 2016.

[16] 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, Aug. 2016.

[17] B. Kimy et al., "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *Proc. IEEE Mil. Commun. Conf*, Nov. 2013, pp. 1278–1283.

[18] M. Zeng, G. I. Tsiropoulos, O. A. Dobre, and M. H. Ahmed, "Power allocation for cognitive radio networks employing non-orthogonal multiple access," in *Proc. IEEE Globecom*, Washington DC, USA, Dec. 2016.

[19] M. Zeng, G. I. Tsiropoulos, A. Yadav, O. A. Dobre, and M. H. Ahmed, "A two-phase power allocation scheme for CRNs employing NOMA," in *Proc. IEEE Globecom*, Singapore, Singapore, Dec. 2017, pp. 1–6.

[20] S. Shi, L. Yang, and H. Zhu, "Outage balancing in downlink nonorthogonal multiple access with statistical channel state information," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4718–4731, Jul. 2016.

[21] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for MC-NOMA systems," in *Proc. IEEE Globecom*, Washington, DC, USA, Dec 2016, pp. 1–6.

[22] Z. Wei, D. W. K. Ng, and J. Yuan, "Power-efficient resource allocation for MC-NOMA with statistical channel state information," in *Proc. IEEE Globecom*, Washington, DC, USA, Dec. 2016, pp. 1–7.

[23] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.

[24] Y. Zhang, H. M. Wang, T. X. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2852–2857, Mar. 2017.

[25] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation for uplink NOMA," in *Proc. IEEE Globecom*, Abu Dhabi, UAE, Dec. 2018, pp. 1–6.

[26] ——, "A fair individual rate comparison between MIMO-NOMA and MIMO-OMA," in *Proc. IEEE Globecom Wkshps*, Singapore, Dec 2017, pp. 1–5.

[27] ——, "Energy-efficient power allocation for hybrid multiple access systems," in *Proc. IEEE ICC Wkshps*, Kansas City, MO, USA, May 2018, pp. 1–5.

[28] M. Zeng, N. P. Nguyen, O. A. Dobre, Z. Ding, and H. V. Poor, "Spectral and energy efficient resource allocation for multi-carrier uplink NOMA systems," *IEEE Trans. Veh. Technol.*, submitted.

[29] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Trans. Commun.*, vol. 98, no. 3, pp. 403–414, Mar. 2015.

[30] J. Choi, "On the power allocation for mimo-noma systems with layered transmissions," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3226–3237, May 2016.

[31] Z. Chen, Z. Ding, P. Xu, and X. Dai, "Optimal precoding for a QoS optimization problem in two-user MISO-NOMA downlink," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1263–1266, Jun. 2016.

[32] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, "A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems," *IEEE Trans. Signal Processing*, vol. 64, no. 1, pp. 76–88, Jan. 2016.

[33] K. Higuchi and Y. Kishiyama, "Non-orthogonal access with random beamforming and intra-beam SIC for cellular MIMO downlink," in *Proc. IEEE VTC Fall*, Sep. 2013, pp. 1–5.

[34] N. Yang, L. Wang, G. Geraci, M. Elkashlan, J. Yuan, and M. D. Renzo, "Safeguarding 5G wireless communication networks using physical layer security," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 20–27, Apr. 2015.

[35] A. D. Wyner, "The wire-tap channel," *Bell Syst. Tech. J.*, vol. 54, no. 8, pp. 1355–1387, Jan. 1975.

[36] Y.-S. Shiu, S. Chang, H.-C. Wu, S. Huang, and H.-H. Chen, "Physical layer security in wireless networks: A tutorial," *IEEE Wireless Commun.*, vol. 18, no. 2, pp. 66–74, Apr. 2011.

[37] J. Chen, L. Yang, and M.-S. Alouini, "Physical layer security for cooperative NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4645–4649, May 2018.

[38] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and L. Hanzo, "Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1656 – 1672, Mar. 2017.

[39] Y. Zhang, H.-M. Wang, Q. Yang, and Z. Ding, "Secrecy sum rate maximization in non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 930–933, May 2016.

[40] B. He, A. Liu, N. Yang, and V. K. N. Lau, "On the design of secure non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2196–2206, Oct. 2017.

Chapter 2

[1] 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 Veh. Technol. Conf.*, Dresden, Germany, Jun. 2013, pp. 1–5.

[2] L. Dai, B. Wang, Y. Yuan, S. Han, C. l. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[3] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surv. Tuts.*, vol. pp, no. 99, pp. 1–1, Oct. 2016.

[4] M. Zeng, G. I. Tsiropoulos, O. A. Dobre, and M. H. Ahmed, "Power allocation for cognitive radio networks employing non-orthogonal multiple access," in *Proc. IEEE Global Telecommun. Conf.*, Washington DC, USA, Dec. 2016.

[5] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5G wireless networks," in *Proc. Int. Symp. Wireless Commun. Systems*, Barcelona, Spain, Aug. 2014, pp. 781–785.

[6] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag. – Feature Topic on LTE Evolution,* to appear. *arXiv preprint arXiv:1511.08610*, 2015.

[7] S. M. R. Islam, M. Zeng, and O. A. Dobre, "Noma in 5G systems: Exciting possibilities for enhancing spectral efficiency," *IEEE 5G Tech. Focus,* to appear, 2017.

[8] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[9] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication.* Cambridge, UK: Cambridge University Press, 2005.

[10] 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, Dec. 2014.

[11] Q. Sun, S. Han, I. Chin-Lin, and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 4, pp. 405–408, Dec. 2015.

[12] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[13] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, Jun. 2016.

[14] Y. Liu, G. Pan, H. Zhang, and M. Song, "On the capacity comparison between MIMO-NOMA and MIMO-OMA," *IEEE Access*, vol. 4, no. 6, pp. 2123–2129, Jul. 2016.

[15] M. Zeng, Y. Animesh, 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.,* DOI: 10.1109/LWC.2017.2712149.

[16] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7244–7257, Nov. 2016.

[17] Z. Ding and H. V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 629–633, May 2016.

Chapter 2

[1] S. M. R. Islam, M. Zeng, and O. A. Dobre, "NOMA in 5G systems: Exciting possibilities for enhancing spectral efficiency," *IEEE 5G Tech. Focus,* vol. 1, no. 2, May 2017. [Online]. Available: http://5g.ieee.org/tech-focus.

[2] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surv. Tuts.*, vol. pp, no. 99, pp. 1–1, Oct. 2016.

[3] L. Dai, B. Wang, Y. Yuan, S. Han, C. L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[4] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[5] L. Song, Y. Li, Z. Ding, and H. V. Poor, "Resource management in non-orthogonal multiple access networks for 5G and beyond," *IEEE Network*, vol. 31, no. 4, pp. 8–14, Jul. 2017.

[6] 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 Veh. Technol. Conf.*, Dresden, Germany, Jun. 2013, pp. 1–5.

[7] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication.* Cambridge, UK: Cambridge University Press, 2005.

[8] 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, Dec. 2014.

[9] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, Jun. 2016.

[10] 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, Aug. 2017.

[11] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[12] 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. Select. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.

[13] B. Wang, L. Dai, Z. Wang, N. Ge, and S. Zhou, "Spectrum and energy efficient beamspace MIMO-NOMA for millimeter-wave communications using lens antenna array," *IEEE J. Select. Areas Commun.*, to appear, 2017.

[14] V. Nguyen, H. Tuan, T. Duong, H.V., Poor, and O. Shin, "Precoder design for signal superposition in MIMO-NOMA multicell networks," *IEEE J. Select. Areas Commun.*, vol. PP, no. 99, pp. 1–1, Jul. 2017.

[15] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722–3732, Sep. 2016.

[16] Y. Zhang, H. M. Wang, T. X. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2852–2857, Mar. 2017.

[17] W. M. Hao, et al., "Energy-efficient power allocation in millimeter wave massive mimo with non-orthogonal multiple access," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 782–785, Dec 2017.

[18] M. Zeng, G. I. Tsiropoulos, O. A. Dobre, and M. H. Ahmed, "Power allocation for cognitive radio networks employing non-orthogonal multiple access," in *Proc. IEEE Global Commun. Conf.*, Washington DC, USA, Dec. 2016.

[19] A. Zappone, P. Lin, and E. Jorswieck, "Energy efficiency in secure multi-antenna systems," *IEEE Trans. Signal Process.,* submitted for publication. [Online]. Available: http://arxiv.org/abs/1505.02385.

Chapter 4

[1] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation for uplink NOMA," in *Proc IEEE Globecom*, Abu Dhabi, UAE, Dec. 2018, pp. 1–6.

[2] S. M. R. Islam, M. Zeng, O. A. Dobre, and K. Kwak, "Resource allocation for downlink NOMA systems: Key techniques and open issues," *IEEE Wireless Commun. Mag.*, vol. 25, no. 2, pp. 40–47, Apr. 2018.

[3] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[4] S. M. R. Islam, M. Zeng, and O. A. Dobre, "NOMA in 5G systems: Exciting possibilities for enhancing spectral efficiency," *IEEE 5G Tech. Focus,* vol. 1, no. 2, May 2017. [Online]. Available: 307 http://5g.ieee.org/tech-focus.

[5] L. Dai, B. Wang, Y. Yuan, S. Han, C. L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[6] S. M. R. Islam, et al., "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surv. Tuts.*, vol. 19, no. 2, pp. 721–742, Second quarter 2017.

[7] F. Zhou, Y. Wu, R. Q. Hu, Y. Wang, and K. K. Wong, "Energy-efficient NOMA enabled heterogeneous cloud radio access networks," *IEEE Network*, vol. 32, no. 2, pp. 152–160, Mar. 2018.

[8] M. Zeng, G. I. Tsiropoulos, O. A. Dobre, and M. H. Ahmed, "Power allocation for cognitive radio networks employing non-orthogonal multiple access," in *Proc. IEEE Global Commun. Conf.*, Washington DC, USA, Dec. 2016.

[9] 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, Dec. 2014.

[10] 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, Aug. 2017.

[11] ——, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Select. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.

[12] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686–7698, Nov. 2016.

[13] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "A fair individual rate comparison between MIMO-NOMA and MIMO-OMA," in *Proc IEEE Globecom Wkshps*, Singapore, Dec 2017, pp. 1–5.

[14] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722–3732, Sep. 2016.

[15] Y. Zhang, H. M. Wang, T. X. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2852–2857, Mar. 2017.

[16] W. M. Hao et al., "Energy-efficient power allocation in millimeter wave massive MIMO with non-orthogonal multiple access," *IEEE Wireless Commun. Lett.*, vol. PP, no. 99, pp. 1–1, Jun. 2017.

[17] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation for MIMO-NOMA with multiple users in a cluster," *IEEE Access*, vol. 6, pp. 5170–5181, Feb. 2018.

[18] W. Hao, Z. Chu, F. Zhou, S. Yang, G. Sun, and K. Wong, "Green communication for NOMA-based CRAN," *IEEE Internet of Things J.*, pp. 1–1, 2018.

[19] T. Lv, Y. Ma, J. Zeng, and P. T. Mathiopoulos, "Millimeter-wave NOMA transmission in cellular M2M communications for internet of things," *IEEE Internet of Things J.*, vol. 5, no. 3, pp. 1989–2000, Jun. 2018.

[20] Y. Endo, Y. Kishiyama, and K. Higuchi, "Uplink non-orthogonal access with MMSE-SIC in the presence of inter-cell interference," in *Proc IEEE ISWCS*, Aug. 2012, pp. 261–265.

[21] X. Chen, A. Benjebbour, A. Li, and A. Harada, "Multi-user proportional fair scheduling for uplink non-orthogonal multiple access (NOMA)," in *Proc. IEEE VTC*, May. 2014, pp. 1–5.

[22] W. Liu, X. Hou, and L. Chen, "Enhanced uplink non-orthogonal multiple access for 5G and beyond systems," *Front. Inform. Technol. Electron. Eng.*, vol. 19, no. 3, pp. 340–356, Mar. 2018.

[23] Y. Liang, X. Li, J. Zhang, and Z. Ding, "Non-orthogonal random access for 5G networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4817–4831, July 2017.

[24] M. Shirvanimoghaddam, M. Condoluci, M. Dohler, and S. J. Johnson, "On the fundamental limits of random non-orthogonal multiple access in cellular massive IoT," *IEEE J. Select. Areas Commun.*, vol. 35, no. 10, pp. 2238–2252, Oct. 2017.

[25] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X. Xia, "Joint power control and beamforming for uplink non-orthogonal multiple access in 5g millimeter-wave communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6177–6189, Sep. 2018.

[26] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (noma) systems," *IEEE Access*, vol. 4, pp. 6325–6343, Aug. 2016.

[27] D. Zhai and J. Du, "Spectrum efficient resource management for multi-carrier-based NOMA networks: A graph-based method," *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 388–391, Jun. 2018.

[28] T. Lv, Z. Lin, P. Huang, and J. Zeng, "Optimization of the energy-efficient relay-based massive IoT network," *IEEE Internet of Things J.*, vol. 5, no. 4, pp. 3043–3058, Aug. 2018.

[29] Z. Yang, W. Xu, H. Xu, J. Shi, and M. Chen, "Energy efficient non-orthogonal multiple access for machine-to-machine communications," *IEEE Commun. Lett.*, vol. 21, no. 4, pp. 817–820, Apr. 2017.

[30] M. Zeng, W. Hao, O. A. Dobre, and V. Poor, "Energy-efficient power allocation in uplink mmwave massive MIMO with NOMA," *IEEE Trans. Veh. Technol.*, pp. 1–1, 2019.

[31] E. Bodine-Baron, C. Lee, A. Chong, B. Hassibi, and A. Wierman, "Peer effects and stability in matching markets," in *Proc. 4th Symp. Algorithmic Game Theory (SAGT)*, Amalfi, Italy, Oct. 2011, pp. 117–129.

[32] A. Zappone, P. Lin, and E. Jorswieck, "Energy efficiency in secure multi-antenna systems," *IEEE Trans. Signal Process.,* submitted for publication. [Online]. Available: http://arxiv.org/abs/1505.02385.

[33] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics (NRL)*, vol. 1-2, pp. 83–97, 1955.

Chapter 5

[1] V. W. S. Wong et al., *Key Technologies for 5G Wireless Systems.* Cambridge, UK: Cambridge University Press, 2017.

[2] S. M. R. Islam, M. Zeng, and O. A. Dobre, "NOMA in 5G systems: Exciting possibilities for enhancing spectral efficiency," *IEEE 5G Tech. Focus,* vol. 1, no. 2, May 2017. [Online]. Available: 307 http://5g.ieee.org/tech-focus.

[3] S. M. R. Islam, M. Zeng, O. A. Dobre, and K. Kwak, "Resource allocation for downlink noma systems: Key techniques and open issues," *IEEE Wireless Commun. Mag.*, vol. 25, no. 2, pp. 40–47, April 2018.

[4] S. M. R. Islam et al., "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surv. Tuts.*, vol. 19, no. 2, pp. 721–742, Second quarter 2017.

[5] M. Zeng, G. I. Tsiropoulos, O. A. Dobre, and M. H. Ahmed, "Power allocation for cognitive radio networks employing non-orthogonal multiple access," in *Proc. IEEE Globecom*, Washington DC, USA, Dec. 2016.

[6] Z. Wei, D. W. K. Ng, J. Yuan, and H. Wang, "Optimal resource allocation for power-efficient mc-noma with imperfect channel state information," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3944–3961, Sep. 2017.

[7] 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, Oct. 2017.

[8] ——, "On the sum rate of MIMO-NOMA and MIMO-OMA systems," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 534–537, Aug. 2017.

[9] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[10] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "A fair individual rate comparison between MIMO-NOMA and MIMO-OMA," in *Proc. IEEE Globecom Wkshps*, Singapore, Dec 2017, pp. 1–5.

[11] N. Yang, L. Wang, G. Geraci, M. Elkashlan, J. Yuan, and M. D. Renzo, "Safeguarding 5G wireless communication networks using physical layer security," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 20–27, Apr. 2015.

[12] A. D. Wyner, "The wire-tap channel," *Bell Syst. Tech. J.*, vol. 54, no. 8, pp. 1355–1387, Jan. 1975.

[13] Y.-S. Shiu, S. Chang, H.-C. Wu, S. Huang, and H.-H. Chen, "Physical layer security in wireless networks: A tutorial," *IEEE Wireless Commun.*, vol. 18, no. 2, pp. 66–74, Apr. 2011.

[14] J. Chen, L. Yang, and M.-S. Alouini, "Physical layer security for cooperative NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4645–4649, May 2018.

[15] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and L. Hanzo, "Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1656 – 1672, Mar. 2017.

[16] Y. Zhang, H.-M. Wang, Q. Yang, and Z. Ding, "Secrecy sum rate maximization in non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 930–933, May 2016.

[17] B. He, A. Liu, N. Yang, and V. K. N. Lau, "On the design of secure non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2196–2206, Oct. 2017.

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

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

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

[21] W. Hao, M. Zeng, Z. Chu, S. Yang, and G. Sun, "Energy-efficient resource allocation for mmWave massive MIMO HetNets with wireless backhaul," *IEEE Access*, vol. 6, pp. 2457–2471, Feb. 2018.

[22] W. Hao, M. Zeng, Z. Chu, and S. Yang, "Energy-efficient power allocation in millimeter wave massive MIMO with non-orthogonal multiple access," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 782–785, Dec. 2017.

[23] M. Zeng, W. Hao, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation in uplink mmWave massive MIMO with NOMA," *IEEE Trans. Veh. Technol.*, pp. 1–1, 2019.

[24] J. Ma, C. Liang, C. Xu, and L. Ping, "On orthogonal and superimposed pilot schemes in massive MIMO NOMA systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2696 – 2707, Dec. 2017.

[25] X. Chen, Z. Zhang, C. Zhong, D. W. K. Ng, and R. Jia, "Exploiting inter-user interference for secure massive non-orthogonal multiple access," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 4, pp. 788–801, Apr. 2018.

[26] J. Zhu, R. Schober, and V. K. Bhargava, "Linear precoding of data and artificial noise in secure massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 2245–2261, Mar. 2016.

[27] N.-P. Nguyen, H. Q. Ngo, T. Q. Duong, H. D. Tuan, and K. Tourki, "Secure massive MIMO with the artificial noise-aided downlink training," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 4, pp. 802 – 816, Apr. 2018.

[28] Y.-Y. Zhang, J.-K. Zhang, and H.-Y. Yu, "Physically securing energy-based massive MIMO MAC via joint alignment of multi-user constellations and artificial noise," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 4, pp. 829 – 844, Apr. 2018.

[29] N.-P. Nguyen, H. Q. Ngo, T. Q. Duong, H. D. Tuan, and D. B. da Costa, "Full-duplex cyber-weapon with massive arrays," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5544 – 5558, Aug. 2017.

[30] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. San Diego, CA: Academic press, 2007.

[31] H. Tabassum, E. Hossain, and J. Hossain, "Modeling and analysis of uplink non-orthogonal multiple access in large-scale cellular networks using poisson cluster processes," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3555–3570, Aug 2017.

[32] H. Sun, B. Xie, R. Q. Hu, and G. Wu, "Non-orthogonal multiple access with sic error propagation in downlink wireless mimo networks," in *Proc. IEEE VTC*, Sep. 2016, pp. 1–5.

[33] X. Chen, Z. Zhang, C. Zhong, R. Jia, and D. W. K. Ng, "Fully non-orthogonal communication for massive access," *IEEE Trans. Commun.*, vol. 66, no. 4, pp. 1717–1731, Apr. 2018.

[34] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation for MIMO-NOMA with multiple users in a cluster," *IEEE Access*, vol. 6, pp. 5170–5181, 2018.

[35] ——, "Energy-efficient power allocation for hybrid multiple access systems," in *Proc. IEEE ICC Wkshps*, Kansas City, MO, USA, May 2018, pp. 1–5.

[36] W. Hao, Z. Chu, F. Zhou, S. Yang, G. Sun, and K. Wong, "Green communication for NOMA-based CRAN," *IEEE Internet of Things J.*, pp. 1–1, 2018.

[37] M. Zeng, A. Yadav, O. A. Dobre, and H. V. Poor, "Energy-efficient power allocation for uplink NOMA," in *Proc. IEEE Globecom*, Abu Dhabi, UAE, Dec. 2018, pp. 1–6.

[38] ——, "Energy-efficient joint User-RB association and power allocation for uplink hybrid NOMA-OMA," *IEEE Internet of Things J.*, pp. 1–1, 2019.

[39] H. H. Kha, H. D. Tuan, and H. H. Nguyen, "Fast global optimal power allocation in wireless networks by local d.c. programming," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 510–515, Feb. 2012.

[40] N. Vucic, S. Shi, and M. Schubert, "DC programming approach for resource allocation in wireless networks," in *Proc. Int. Symp. Modeling Optimization Mobile, Ad Hoc Wireless Netw.*, May 2010, pp. 380–386.

[41] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21." *Available: http://cvxr.com/cvx*, Dec. 2010.

[42] W. Dinkelbach, "On nonlinear fractional programming," *Manag. Sci.*, vol. 13, no. 7, pp. 3492–498, Mar. 1967.