8-2019

Spectral, Energy and Computation Efficiency in Future 5G Wireless Networks

Haijian Sun
Utah State University

SPECTRAL, ENERGY AND COMPUTATION EFFICIENCY IN FUTURE 5G WIRELESS NETWORKS

by

Haijian Sun

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

of

DOCTOR OF PHILOSOPHY

in

Electrical Engineering

Approved:

Rose Qingyang Hu, Ph.D. Major Professor	Bedri Cetiner, Ph.D. Committee Member
Ryan Davidson, Ph.D. Committee Member	Don Cripps, Ph.D. Committee Member
Haitao Wang, Ph.D. Committee Member	Richard S. Inouye, Ph.D. Vice Provost for Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2019

ABSTRACT

Spectral, Energy and Computation Efficiency in Future 5G Wireless Networks

by

Haijian Sun, Doctor of Philosophy

Utah State University, 2019

Major Professor: Rose Qingyang Hu, Ph.D.
Department: Electrical and Computer Engineering

The past few decades have witnessed the prosperity of wireless communication technologies, from first generation (1G) in the 1980s that focused on short-range voice to 4G in the 2010s that supported high-speed mobile Internet. Driven by an ever-increasing demand on today's applications, with users requesting for high-definition videos, Internet of things (IoT) and autonomous driving, wireless technology continues to evolve. The frontier 5G, expected to debut in 2020, aims to provide 1,000 times more aggregate data rate, 100 times more connected devices, and 10 times lower power consumption. These ambitious goals create significant challenges for spectral, energy, and latency requirements.

In this dissertation, we systematically study spectral, energy and computation efficiency in 5G, the potentials and associated challenges. Specifically, we introduce non-orthogonal multiple access (NOMA), a promising access technique, and mobile edge computing (MEC), an emerging communication scheme. NOMA is a new multiple access technique that allows more users to share the same time/frequency resource, which can further improve the system spectral efficiency, as well as allow more devices to connect. MEC, on the other hand, is a new scheme that permits processing units in the network edge, which can significantly reduce system latency. This dissertation has applied NOMA in various communication scenarios and studied their performance, as well as proposed a new evaluation in MEC. We

have conducted extensive simulations to verify the feasibility of our proposed algorithms and new schemes. Lastly, we applied both NOMA and MEC into a practical wearable communication architecture to show that this scheme can meet the diverse communication needs of wearable devices.

(159 pages)

PUBLIC ABSTRACT

Spectral, Energy and Computation Efficiency in Future 5G Wireless Networks

Haijian Sun

Wireless technology has revolutionized the way people communicate. From first generation, or 1G, in the 1980s to current, largely deployed 4G in the 2010s, we have witnessed not only a technological leap, but also the reformation of associated applications. It is expected that 5G will become commercially available in 2020. 5G is driven by ever-increasing demands for high mobile traffic, low transmission delay, and massive numbers of connected devices. Today, with the popularity of smart phones, intelligent appliances, autonomous cars, and tablets, communication demands are higher than ever, especially when it comes to low-cost and easy-access solutions.

Existing communication architecture cannot fulfill 5G's needs. For example, 5G requires connection speeds up to 1,000 times faster than current technology can provide. Also, from transmitter side to receiver side, 5G delays should be less than 1 ms , while 4G targets a 5 ms delay speed. To meet these requirements, 5G will apply several disruptive techniques. We focus on two of them: new radio and new scheme. As for the former, we study the non-orthogonal multiple access (NOMA) and as for the latter, we use mobile edge computing (MEC).

Traditional communication systems allow users to communicate alternatively, which clearly avoids inter-user interference, but also caps the connection speed. NOMA, on the other hand, allows multiple users to transmit simultaneously. While NOMA will inevitably cause excessive interference, we prove such interference can be mitigated by an advanced receiver side technique. NOMA has existed on the research frontier since 2013. Since that time, both academics and industry professionals have extensively studied its performance. In this dissertation, our contribution is to incorporate NOMA with several potential

schemes, such as relay, IoT, and cognitive radio networks. Furthermore, we reviewed various limitations on NOMA and proposed a more practical model.

In the second part, MEC is considered. MEC is a transformation from the previous cloud computing system. In particular, MEC leverages powerful devices nearby and instead of sending information to distant cloud servers, the transmission occurs in closer range, which can effectively reduce communication delay. In this work, we have proposed a new evaluation metric for MEC which can more effectively leverage the trade-off between the amount of computation and the energy consumed thereby.

A practical communication system for wearable devices is proposed in the last part, which combines all the techniques discussed above. The challenges for wearable communication are inherent in its diverse needs, as some devices may require low speed but high reliability (factory sensors), while others may need low delay (medical devices). We have addressed these challenges and validated our findings through simulations.

To my family.

ACKNOWLEDGMENTS

Firstly of all, I would like to express my sincere appreciation to my advisor, Prof. Rose Q. Hu. She provides tremendous helps throughout my entire Ph.D. career, not only with my researches, but also as an inspirational role model. Prof. Hu guides me how to perform excellent research, how to collaborate with other people, and how to instruct in classes. Most importantly, she teaches me how to be professional. These valuable experiences will surely make myself a better researcher and advisor in the future.

I also thank Prof. Bedri Cetiner, Prof. Ryan Davidson, Prof. Don Cripps, and Prof. Haitao Wang for their service as my committee members, I received valuable comments and suggestions for the proposal and dissertation. Additionally, Prof. Todd Moon, Prof. Jacob Gunther, and Prof. Chris Winstead for their classes and valuable discussions.

Next, I am grateful to Prof. Yi Qian, Dr. Geng Wu, Prof. Feng Ye, and Dr. Fuhui Zhou, who helped me with many research projects and topics. They always provided me with insightful ideas and discussions.

For my colleagues and friends, together with those listed above, Dr. Yiran Xu, Dr. Xue Chen, Dr. Zekun Zhang, Xuan Xie, Dr. Tao Xu, Dr. Jingru Zhang, Dr. Min Xian, Dr. Le Thanh Tan, Qun Wang, Shakil Ahmed, they helped me with various research discussions and other aspects, which makes my life here more colorful.

Furthermore, I would like to thank Dr. Perry Wang, Dr. Milutin Pajovic, Dr. Toshiaki Koike-Akino, and Dr. Philip Orlik for their support and help during my summer internship at Mitsubishi Electric Research Laboratories (MERL) in Cambridge, MA.

Lastly, my mom, my brother and grandma are always supporting me, sharing my happiness and tough times. My wife, Fei Lu, accompanied me through those Ph.D. days with her encouragement and love.

This work has been supported by National Science Foundation under grants ECCS-1308006, NeTS-1423348, and EARS-1547312.

Haijian Sun

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS

D2D	Device-to-device Communication
MIMO	Muitiple-input-multiple-output
HetNet	Heterogeneous Network
BS	Base Station
NOMA	Non-orthogonal Multiple Access
MEC	Mobile Edge Computing
SE	Spectral Efficiency
SS	Superimposed Signal
EE	Energy Efficiency
OMA	Orthogonal Multiple Access
MUD	Multiuser Detection
CR	Cognitive Radio
GA	Genetic Algorithm
CSI	Channel State Information
UE	User Equipment
DF	Decode-and-forward (in relay systems)
EP	Error Propagation
SWIPT	Simultaneous Wireless Information and Power Transfer
mmWave	Millimeter Wave
PU	Primary User
SU	Secondary User
EH	Energy Harvest
CE	Computation Efficiency
ZF	Zero-forcing
SVD	Singular Value Decomposition
LTE	Long Term Evolution
KKT	Karush-Kuhn-Tucker

CHAPTER 1

INTRODUCTION

The past few decades have witnessed unprecedented growth in wireless communication technology. Starting in the 1980s, when the first mobile phone was released by Motorola, technological advancement has exploded in almost every decade, from first generation (1G) to 4G communication capabilities. The invention of smart devices, such as phones, tablets and home appliances, is the driving force for ever-increasing mobile traffic today. It is anticipated that mobile traffic will increase 10-fold between 2014 and 2019 globally. Mobile data traffic is expected to grow much faster than fixed IP traffic in the upcoming years [1]. Because current wireless communication systems, such as LTE, have been pushed to their theoretical capacity limits, different air interface and radio access technologies, including heterogeneous network (HetNet) [2] [3], multiuser multi-input multi-output (MU-MIMO) [4] and device-to-device (D2D) communication [5] have become potential paradigms to fill the gap between end user demands and the capacity that current air interface can provide.

In their pioneering work, Andrews *et al.* evaluated the requirements for 5G. In short, a 5G wireless communication system must provide 1,000 times more aggregate data improvement than 4G, support for as low as 1 *ms* round-trip latencies, a 10 time longer battery life for low-power devices, and the necessary foundation to support 10,000 or more low-rate devices in a single macro cell [6]. This transformation from 4G to 5G cannot be simply fulfilled by extensions of current technologies. In [7], several promising techniques are introduced, notably, 1) More bandwidth. Currently commercial cellular systems use frequencies below 3 GHz (sub-3GHz); in fact, there is abundant bandwidth in the millimeter wave (mmWave) band. In 28 GHz and above, for example, bandwidth is available that previously has not been applied in cellular networks. 2) More antennas. Higher frequency results in smaller form factor of large antenna arrays. The signal processing techniques in terms of massive MIMO and transceiver design also improved significantly. 3) New radios (NR). The physical

layer in 5G will change dramatically, specifically the 5G NR, which includes a new multiple access technology, a new air interface, and a combination of several existing techniques. 4) New schemes. It is expected that ultra-dense networks (UDN) will be heavily deployed. The density of small base stations (BSs), such as micro BS, femto cell and pico cells, will be much higher than that of 4G. But they share similarities in terms of deploying BSs with different powers to provide seamless coverage, as well as performance improvements from short-range communications.

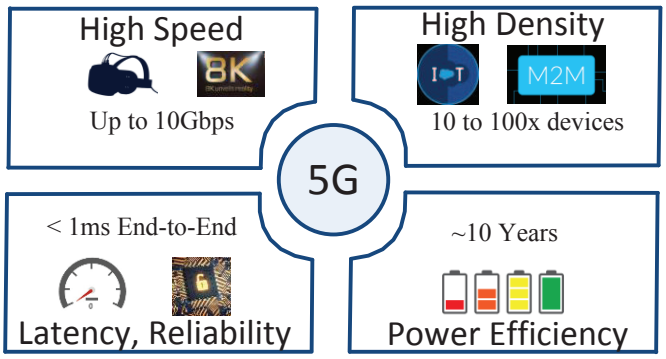

Fig. 1.1: Four main goals for 5G

Though the promising technologies described above are able to deliver on the ambitious goals of 5G, they ultimately encounter some challenges. First of all, mmWave signals are notorious for weak penetration and easy blockage, hence the transmission characteristics are big concerns. Moreover, studies also showed mmWave signals have high attenuation when faced with atmospheric gases, rain, concrete structures, glasses and even foliage. Secondly, from a transceiver design perspective, higher frequency signals impose challenges with regard to circuit design, materials, and heating issues. Nyquist theorem sets the lower boundary for sampling rate in communication systems. With wide bandwidth in an mmWave spectrum, a sampling rate can reach up to 10 Gbit/s, but high-speed circuit design becomes very difficult. It has also been reported that the energy efficiency for components (power amplifier, analog-to-digital converter, digital-to-analog converter) in high frequency is low, registering around 10%. Furthermore, such low efficiency in these components brings with it thermal issues

in hand-held devices. Thirdly, with mmWave band, performance gain largely comes from antenna array systems. Current design can incorporate hundreds of antenna elements in a small area. Even though this can facilitate beamforming, a process which generates narrow but stronger signals in a desired direction, the overhead for channel estimation and precoding is usually too high. Lastly, in UDN networks, challenges in mobility management, interference management and heterogeneity interfere with the nature of the devices.

Recently, several emerging technologies have aimed to deliver the goal of 5G, and thereby solve the challenges above. Specifically, in this dissertation, we consider non-orthogonal multiple access (NOMA), mobile edge computing (MEC), and a new communication architecture for wearable and Internet of Things (IoT) devices. We conducted preliminary research meant to address the goals above. Specifically, NOMA has the capability to improve data rate and support more devices simultaneously, while schemes for wearable communications and MEC can help with power consumption and latency. Below, we briefly introduce each enabling technique.

1.1 NOMA

Initially proposed by NTT DOCOMO as an enhancement for LTE-advanced (LTE-A) in 2013 [28], NOMA has been recognized as one of the most promising 5G techniques, due to its capability of providing high spectral efficiency (SE) and massive connectivity. The basic principle of NOMA is that on the transmitter side, multiple signals are added up with different powers, forming a superimposed signal (SS). On the receiver side, to guarantee a weak user's quality of service (QoS), successive interference cancellation (SIC) is used to retrieve each user's signal sequentially from the SS [8]. Specifically, a user can decode the strongest signal by treating other signals as interference. If the decoded signal is its own data, SIC stops. Otherwise, the receiver will subtract the decoded signal from the SS and continue to decode the next strongest signal. Note that SS with SIC is not new; within information theory, this duo already exists as a capacity-achieving technique in uplink communication. However, within NOMA, the weak user will have stronger power, which is not information-theoretic optimal. Still, since its design philosophy may be combined with

diverse transceivers, it has drawn tremendous attention in multiple-antenna systems [9] and in downlink and uplink multi-cell networks [11]. In contrast to classic orthogonal multiple access (OMA), NOMA provides simultaneous access to multiple users at the same time and on the same frequency band. One method used to accomplish such a feat is by using power-domain multiplexing. It has been shown that NOMA is capable of achieving a higher SE and energy efficiency (EE) than OMA, such as OFDMA, TDMA and FDMA [8]- [11].

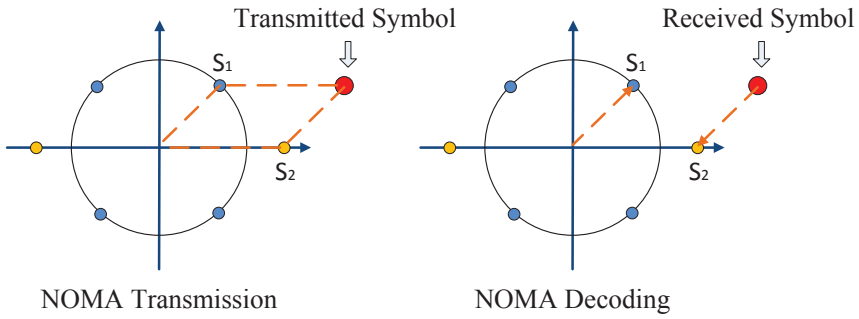

Fig. 1.2: NOMA principles: transmission and decoding stage

Fig. 1.2 shows the basic principle of NOMA with regards to data encoding and decoding. S_1 and S_2 are the symbols used to denote user 1 and user 2, respectively. We also assume user 1 has a better channel than user 2. On the transmitter side, BPSK and QPSK modulation are applied for two users. Clearly, the average symbol power of S_2 is larger to compensate for the unfavorable channel. Actual transmitted symbol is simply the addition of these two symbols. On the receiver side, the symbol with highest power will be decoded firstin this example, S_2. Since the received symbol is on the right side of the y-axis, for BPSK, it will be decoded as S_2 and will be removed from the composite signal, which only has S_1 left. Note that the symbols can use the same modulation scheme as long as they have different power. Furthermore, in this dissertation, we do not consider any specific modulation; rather, we apply the Gaussian coding and perform analysis based on information-theoretic perspective.

The disadvantage of NOMA, however, lies in the following aspects. Firstly, NOMA requires a more complicated receiver structure to perform SIC, hence the cost will be higher

and receiver architecture must also be changed accordingly. Secondly, during the SIC procedure, one user must decode signal from others, which will cause security and privacy concerns. Lastly, depending upon implementation, NOMAs successive decoding will impose certain delays for users.

Starting with the 3rd Generation Partnership Project (3GPP) LTE Release-13, NOMA, as one of the techniques in multi-user superposed transmission (MUST), became part of the standardization. In 2017, with LTE Release-14, 15 MUST schemes have been proposed for the uplink NR. Additionally, NOMA has attracted extensive attention within the industry. NTT DoCoMo and MediaTek collaborated to field test NOMA in Nov. 2017. Using a simple scenario of one base station and three users, they were able to achieve 2.3 times the spectral efficiency available with current technology [1].

Despite its shortcomings, we have applied NOMA in many schemes and systematically studied its performance for example, we have applied NOMA with D2D, MIMO, relay networks and cognitive radio. More importantly, we reviewed the fundamental principles of NOMA and pointed out the error propagation phenomenon. Furthermore, we considered the channel imperfection and its impact on NOMA performance.

1.2 MEC

Due to size, battery and cost limitations, mobile devices can experience performance bottleneck when computation-intensive tasks are added. More than a decade ago, people solved this problem by introducing the concept of cloud computing, whereby mobile devices did not perform large-scale computation locally; instead, they sent such tasks to remote servers for faster and more secure processing, storage and sharing. The centralized nature of cloud-based computing reduced the expenditure cost while providing an easier deployment process. However, cloud servers are often located in remote areas, which inevitably results in longer end-to-end transmission and greater processing delays [12].

MEC is a new alternative paradigm for upcoming 5G systems. Instead of transmitting data to remote cloud-based servers for processing, MEC provides certain computation ca-

[1] MediaTek Newsroom, Nov. 2017

pacities locally even within the base stations of the wireless cellular networks [82]. This paradigm shift can effectively reduce long backhaul latency and energy consumption, as well as provide support for a more flexible infrastructure in a cost-effective way [86]. MEC has attracted extensive research interest recently, not only on the architectural level, but also with respect to specific tasks such as cooperative computation offloading [85]- [89]. Computation offloading, which leverages powerful MEC servers in proximity and sends computation-intensive tasks for further processing, is a desirable scheme to overcome the physical limitations of user devices.

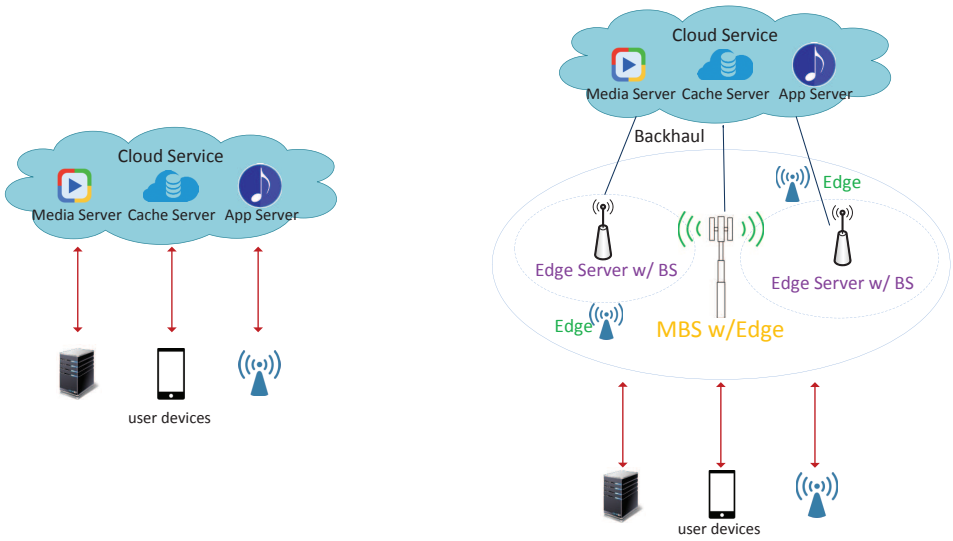

Fig. 1.3: Paradigm shift from cloud computing to mobile edge computing

We see this paradigm shift in a more fundamental way. In the cloud computing era, while the data transmission speed is not high, the bottleneck comes primarily due to computation capacity. With Moore's law still in effect, performance of integrated circuit chips grows exponentially. On the other hand, communication technology causes the speed to increase almost linearly. Since the goal is to reduce processing speed, it will be more beneficial to perform task execution both locally and remotely.

In this dissertation, in order to reduce latency as well as improve system efficiency, we propose a joint processing scheme in which the total task can be divided arbitrarily into two

parts, one for local computing and the other for offloading. To cope with the ever-increasing pressure to ensure energy efficiency, we evaluate the system performance by a new metric, computation efficiency (CE). Computation efficiency is defined as the total number of bits computed within the total energy consumption. The objective is to maximize each user's CE within time constraints (users should finish their task before a required time), energy constraint (each user is powered by battery; hence the total energy should fall below a specific threshold) and task constraint (each user should finish a minimum number of data bits). Later, we show that CE is a more appropriate method in terms of finding a balance of more tasks and less energy.

1.3 Communication architecture for wearables

Recent years have witnessed unprecedented growth in wearable devices, owing to swift advances in chip design, computing, sensing and communications technologies. A wearable device, or simply a wearable, refers to a device that can be worn on the body. While wearable devices are not new, the past few years have seen a surge in their large-scale use and popularity. This rapid rise in popularity was spurred, in part, by technological innovation. Emerging system on chip (SoC) and system in package (SiP) have scaled down the printed circuit board (PCB) size, decreased power consumption, and most importantly, made it possible to design wearables in a variety of desired shapes.

Wearable devices provide easier access to information and convenience for their users. They have varying form factors, from low-end health and fitness trackers to high-end virtual reality (VR) devices, augmented reality (AR) helmets and smart watches. These devices can collect data on heart rates, steps, locations, surrounding buildings, sleeping cycles and even brain waves. Yet computing limitations continue to hinder wearables ability to process data locally. As a result, most send their collected data to other powerful devices or to the clouds. This necessary communication plays a key role in wearable devices and is the main emphasis in this dissertation. Different applications provided by different wearables may have varying communication requirements. For example, while medical sensors have stringent requirements on latency and reliability, they have a low data rate need. On the

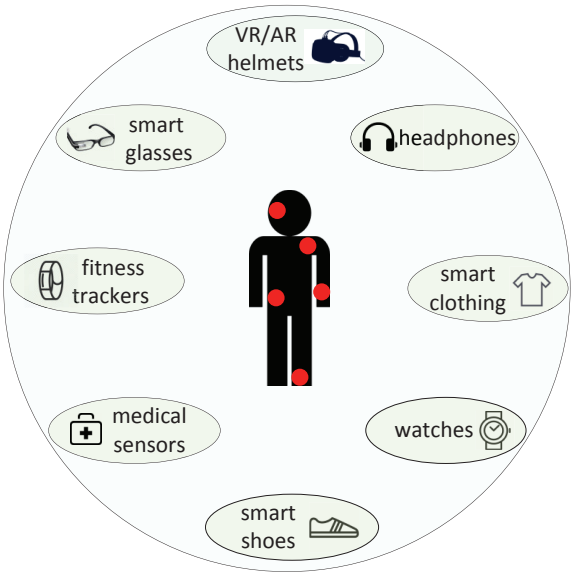

Fig. 1.4: Wearable devices may have varying forms, from small medical sensors to entertainment helmets.

other hand, AR/VR devices need both high throughput and low latency for a better user experience.

Wearable devices may not be able to take full advantage of current communication architecture, due to their potential cost and hardware complexity. On the other hand, wearable devices have succeeded in becoming more and more integrated into everyday activities requiring voice, image, and video inputs. Human beings are generally sensitive to an approximate 100 *ms* audible delay and can catch visual delays of less than 10 *ms*. Furthermore, cell phones and tablets now primarily use touch interaction, a "tactile interaction" that requires a more rigorous delay control, such as 1 *ms*.

1.4 Dissertation Outline

In this dissertation, faced with the several challenges of 5G, we focus on improving spectral, energy and computation efficiency. As mentioned above, we primarily apply NOMA to solve spectral efficiency and the number of connected devices challenges. MEC is studied for CE considerations. The wearable communication architecture combined several techniques in NOMA and in general to help reduce communication latency.

First, in chapter 2, we review the NOMA technique and apply it in a downlink MIMO communication system where a combination of D2D and cellular users exist. In particular, we assume cellular users can be grouped and follow NOMA transmission within each group. We calculate each user's achievable data rate and formulate an optimization problem for maximizing total system throughput. Variables to be optimized are: user grouping indices, power allocation for NOMA users and beamforming vectors to suppress inter-user interference.

Chapter 3 analyzes the SIC process of NOMA and concludes that the assumption of the stronger users always successful decoding is too strong. Hence, we propose the error propagation, which models unsuccessful decoding in the previous stage as residual interference. In light of this, we consider a simple MIMO scenario where both the BS and two users equip with multiple antennas. The objective is to maximize the total data rate of these two users, with the constraint that the weak user's QoS should be satisfied. The variable to be optimized is transmitter beamforming matrix. With error propagation, the problem is non-convex, and we solve it iteratively.

We consider two network settings in Chapter 4: NOMA in relay networks, and NOMA in IoT networks. In the first part, we study two further relay models, namely NOMA cooperative scheme and NOMA TDMA scheme. In NOMA cooperative scheme, the completion of one round of information transmission consists of two time slots. Dirty paper coding (DPC) is used as precoding at relays to eliminate inter-user interference in the second time slot. As a comparison, NOMA TDMA scheme uses three time slots to complete one round of information transmission. In the second part, we consider applying NOMA and D2D relaying in a mmWave-based wireless system that consists of high power base stations and low power IoT devices. The lower power IoT devices do not have external power supplies and have limited battery life. In order to prolong battery life and also to motivate low power IoT devices to help relay signals from others, low power IoT devices can harvest energy from electromagnetic signals. To make the energy harvest model more realistic, a non-linear energy harvesting model is used. The theoretical analysis on outage probability is given

for the proposed scheme and system model. Simulation results validate the accuracy of the analysis.

Chapter 5 analyzes impacts from both error propagation and channel imperfection. In fact, both effects can cause SIC decoding errors. In particular, we consider two channel error models, namely bounded and Gaussian. The former assumes the power of channel estimation error is constrained to a value, and that for higher dimensional data, the errors are in a convex cone. The latter, however, is a more general assumption that errors follow Gaussian distribution. To preserve energy efficiency, we also consider applying simultaneous wireless information and power transfer (SWIPT). Several optimization problems are formulated with differing objectives: either to maximize data rate or to maximize harvested energy.

MEC for joint offloading and local computing is studied in Chapter 6. To do this, we utilize a simple scenario where a BS with MEC server and several low-power devices are deployed. Each device has specific incoming tasks to complete, for which it is impossible to finish the processing timely. Thus, tasks are partitioned into offloading and local computing. CE is proposed to evaluate the system performance, and we desire to have a maximum CE. To achieve this objective, each user's partition ratio, offloading power, local computing CPU frequency, and transmission time is calculated.

Chapter 7 investigates a feasible communication architecture for wearable devices. Different from IoT devices, where the primary concern from a communication perspective is the massive number of devices, wearables have diverse communication needs. We propose a hybrid architecture involving not only cellular 5G, but also evolving wireless local area network (WLAN) techniques. By merging heterogeneous networks and virtual network resource slicing, those diverse needs can be satisfied. Furthermore, we also discuss possible technologies to alleviate power consumption, security and health concerns.

Chapter 8 concludes this dissertation and proposes futureChapter 8 concludes this dissertation and proposes further research directions. research directions.

CHAPTER 2

A NOMA and MIMO Supported Cellular Network with Underlaid D2D Communications

2.1 Introduction

MU-MIMO is one type of MIMO technology for wireless communication, in which multiple spatially distributed users with one or more antennas can be transmitted at the same time and frequency by the base station with multiple antennas, at the cost of signal processing. MU-MIMO can greatly improve the system capacity by exploiting the spatial diversity gain among multiple users. Moreover, we can apply NOMA, in which a transmitter distributes its power among multiple users and then adds up these users' signals, forming a SS to further improve the system spectrum efficiency. Compared with its OMA counterpart, NOMA has a superior performance in terms of spectral efficiency. However, multiuser detection (MUD) is required at the receiver side, which induces a more complex receiver structure and algorithm. In [13], the impact of user pairing on the performance of both fixed power allocation and cognitive radio (CR) inspired NOMA (CR-NOMA) is studied. For the fixed one, NOMA tends to pair users with larger channel gain difference. In [15], a genetic algorithm (GA) based NOMA pairing in the HetNet is presented, GA will help reduce the computation workload.

D2D communication is proposed as another 5G enabler to further improve the system spectral efficiency. D2D users (DUs) can be supported in an underlaid mode, in which they can share the same spectrum with cellular users. Its advantages over traditional communications are multifold: 1) D2D transmitter and receiver are close to each other, which allows lower power transmission. This is important in today's small size battery-driven devices; 2) DUs can share the same spectrum with cellular users (CUs) with careful interference coordination mechanisms. Thus the overall system spectrum efficiency can be enhanced. In [16], DUs underlying a MU-MIMO cellular network is investigated. A more

interesting yet more challenging question is how to jointly consider NOMA, D2D and MU-MIMO altogether. The key part of MU-MIMO is to design a suitable precoding matrix for transmitters based on various objective functions, such as overall system capacity or minimum power consumption. When jointly considering MU-MIMO, NOMA and D2D, a tight coordination among these three mechanisms should be carefully designed so that the overall system performance can be maximized. In [17], NOMA and MU-MIMO are jointly designed to improve the total system throughput. In this paper, we propose a new mechanism that jointly coordinates beamforming based MU-MIMO, NOMA, and D2D communications in a downlink cellular network. By supporting DUs in a NOMA/MU-MIMO cellular network, more complicated interference scenarios arise. To address that, we develop two different precoding schemes. One aims to cancel out the BS to DUs interference while the other one aims to minimize interference among cellular users that coordinate with each other through NOMA and MU-MIMO beamforming. Beamforming is designed together with NOMA pairing and power allocation to significantly improve overall system sum throughput.

2.2 System Model

We consider a downlink MU-MIMO cellular network that jointly supports NOMA, MU-MIMO and underlaying DUs. M CUs are randomly distributed, each equipped with a single antenna. Each BS has N antennas and thus can maximally generate N beamforming vectors. Each beam can support multiple single antenna users by using NOMA, compared with one user in the conventional MU-MIMO system. Furthermore, a total number of P underlaid DU pairs, denoted as DU_1, DU_2, ..., DU_P, are also randomly distributed, to further exploit current spectrum reuse.

For beam n, NOMA allows a set of $\mathbf{\Phi}_n = \{u(n,1), u(n,2), \ldots, u(n,K)\}$ CUs to be scheduled on the same radio resource simultaneously, $K \geq 2$. We use $u(n,k)$ to denote the CU that is served by beam n with NOMA sequence k in that beam. Assume x_n is the transmitted signal in the n-th beam, and according to NOMA, x_n is a superimposed signal

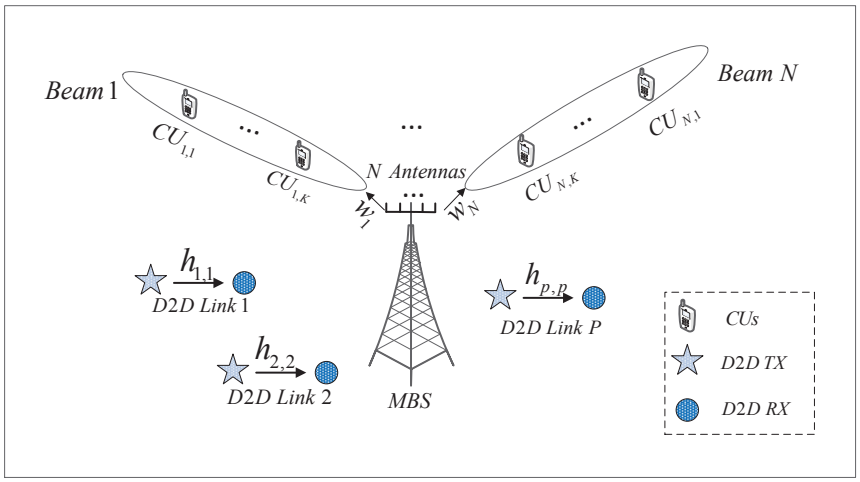

Fig. 2.1: System model

of a total K users in beam n,

$$x_n = \sum_{k=1}^{K} \sqrt{\lambda_{u(n,k)} P_n} s_{u(n,k)}. \tag{2.1}$$

Here $\mathbb{E}(|s_{u(n,k)}|^2) = 1$, $\mathbb{E}(.)$ is the expectation function. $\lambda_{u(n,k)}$ is the fraction of the allocated power to user $u(n,k)$, $\sum_{k=1}^{K} \lambda_{u(n,k)} = 1$. P_n is the total transmitted power for beam n. The total transmission power of a BS is equally partitioned among N beams, i.e., $P_n = \frac{P_{MBS}}{N}$, where P_{MBS} is the total BS transmission power.

At each MBS, a precoding scheme is applied to support MU-MIMO. We denote the precoding matrix as $\mathbf{W}$, which consists of N vectors, i.e.,

$$\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N], \tag{2.2}$$

where $\mathbf{w}_n \in \mathbb{C}^{N\times 1}$ is the beamforming vector of the n-th beam. The received signals at $u(n,k)$ and DU p can be respectively expressed as

$$y_{u(n,k)} = \mathbf{h}_{u(n,k)} \sum_{n=1}^{N} \mathbf{w}_n x_n + \sum_{p=1}^{P} \sqrt{P_D} h_{p,u(n,k)} s_p + n_{u(n,k)} \tag{2.3}$$

$$y_{DU_p} = \sum_{p'=1}^{P} \sqrt{P_D} h_{p',p} s'_p + \mathbf{h}_p \sum_{n=1}^{N} \mathbf{w}_n x_n + n_p, \tag{2.4}$$

where s_p is the transmitted signal of DU p. We also have $\mathbb{E}(|s_p|^2) = 1$. P_D is the transmission power of DUs. $\mathbf{h}_{u(n,k)}$ and $\mathbf{h}_p$ are the channel gains for downlink CU $u(n,k)$ and for DU p, respectively. $h_{p,u(n,k)}$ is the channel gain between DU p and CU $u(n,k)$, and similarly $h_{p',p}$ is the channel gain between the transmitter of DU p' and the receiver of DU p. We assume the channel information is perfectly know at the BS. $n_{u(n,k)}$ and n_p are i.i.d. additive white Gaussian noise at CU $u_(n,k)$ and DU p, respectively. $(n_{u(n,k)}, n_p) \sim \mathcal{CN}(0,1)$.

2.3 NOMA with SIC and Problem Formation

2.3.1 NOMA with SIC

NOMA is a technique that enables multiple users to share the same spectrum resource simultaneously by employing interference cancellation at the receiver. Within a NOMA group, CU with a weaker channel is normally allocated a higher downlink transmission power so that the strongest received signal within that NOMA group corresponds to the CU with the weakest channel gain in that group. The key idea of SIC is that the received SS is decoded in the ascending order of the respective channel gains or in the descending order of the received signal strength, for all the signals that constitute the SS. The receiver decodes the strongest user signal by treating weaker signals in the SS as interference. The decoded signal can be either the desired signal or can be subtracted from the SS. The decoding process will continue until the receiver successfully decodes its own signal [10].

Channel gains for CUs in the same NOMA group in beam n can be sorted as $|\mathbf{h}_{u(n,1)}| \leq |\mathbf{h}_{u(n,2)}| \leq \ldots \leq |\mathbf{h}_{u(n,K)}|$. Since the decoding order follows the ascending order of channel

gains, CU j will decode CU i message, if $i < j$. SIC then removes the decoded message from its observation. CU i treats signals from CUs with index $j > i$ as interference. Assuming perfect interference cancellation, we can rewrite (2.3) as

$$y_{u(n,k)} = \mathbf{h}_{u(n,k)}\mathbf{w}_n\sqrt{\lambda_{u(n,k)}P_n}s_{u(n,k)} + \mathbf{h}_{u(n,k)}\mathbf{w}_n\sum_{k'=1,k'\neq k}^{K}\sqrt{\lambda_{u(n,k')}P_n}s_{u(n,k')}$$
$$+\mathbf{h}_{u(n,k)}\sum_{n'=1,n'\neq n}^{N}\mathbf{w}_{n'}\sum_{k'=1}^{K}\sqrt{\lambda_{u(n',k')}P_{n'}}s_{u(n',k')} + \sum_{p=1}^{P}\sqrt{P_D}h_{p,u(n,k)}s_p + n_{u(n,k)}, \quad (2.5)$$

where the second term on the right side is the interference from users in the same NOMA group. The third term represents inter-beam interference. After applying SIC, the received signal-to-noise-plus-interference-ratio (SINR) $\gamma_{u(n,k)}$ of CU $u(n,k)$ $\gamma_{u(n,k)}$ becomes

$$\gamma_{u(n,k)} = \frac{\lambda_{u(n,k)}P_n|\mathbf{h}_{u(n,k)}\mathbf{w}_n|^2}{I^N_{u(n,k)} + I^U_{u(n,k)} + I^D_{u(n,k)} + \sigma_n^2}, \quad (2.6)$$

where

$$I^N_{u(n,k)} = \sum_{k'=k+1}^{K}\lambda_{u(n,k')}P_n|\mathbf{h}_{u(n,k)}\mathbf{w}_n|^2, \quad (2.7)$$

$$I^U_{u(n,k)} = \sum_{n'=1,n'\neq n}^{N}P_{n'}|\mathbf{h}_{u(n,k)}\mathbf{w}_{n'}|^2, \quad (2.8)$$

$$I^D_{u(n,k)} = \sum_{p=1}^{P}P_D|h_{p,u(n,k)}|^2, \quad (2.9)$$

respectively represent SIC, inter-beam and DU interference to CU $u(n,k)$.

Similarly, SINR γ_{DU_p} of the DU p is expressed as

$$\gamma_{DU_p} = \frac{P_D|h_{p,p}|^2}{\sum_{p'=1,p'\neq p}^{P}P_D|h_{p',p}|^2 + \sum_{n=1}^{N}P_n|\mathbf{h}_p\mathbf{w}_n|^2 + \sigma_n^2}. \quad (2.10)$$

Given SINR, the corresponding user data rate can be calculated as $f(\mathbb{E}\{\gamma\})$ by using Shannon capacity formula,

$$f(\mathbb{E}\{\gamma\}) = \log(1 + \mathbb{E}\{\gamma\}). \tag{2.11}$$

Here we normalize the bandwidth at MBS to 1.

2.3.2 Problem Formation

The design objective is to maximize the total system sum throughput from both CUs and DUs. To this end, we need to determine 1) the NOMA set of each beam, i.e., $\mathbf{\Phi}_n$; 2) the power allocation factor $\lambda_{u(n,k)}$ for each user k in the NOMA set of beam n; and 3) the precoding vector $\mathbf{w}_n$. Therefore, the problem can be formulated as follows.

$$\max_{\Phi_n, \mathbf{w}_n, \lambda_{u(n,k)}} \sum_{n=1}^{N} \sum_{k=1}^{K} f(\mathbb{E}\{\gamma_{u(n,k)}\}) + \sum_{p=1}^{P} f(\mathbb{E}\{\gamma_{DU_p}\}) \tag{2.12}$$

subject to

$$\sum_{k=1}^{K} \lambda_{u(n,k)} = 1,\ n = 1, 2, \ldots, N, \tag{2.13}$$

$$f(\mathbb{E}\{\gamma_{u(n,k)}\}) > R_0,\ \forall k \neq K, \tag{2.14}$$

$$\mathbf{w}_n \in \mathbb{C}^{N \times 1}. \tag{2.15}$$

Constraint (2.13) is the summation of user power in one beam. Constraint (2.14) sets a lower rate limit for users that experience SIC interference in NOMA to ensure good user experience. $\gamma_{u(n,k)}$ and γ_{DU_p} are rates calculated based on (2.6) and (2.10), respectively. The optimization problem is a non-convex problem that needs to determine $\Phi_n, \mathbf{w}_n, \lambda_{u(n,k)}$ jointly. To make this problem feasible to solve, in the next section, we seek a heuristic solution by decomposing the original problem into two sub-problems. We first develop different precoding methods, which aim to suppress either the inter-beam interference among CUs or the interference from CUs to DUs. Based on the precoding matrices, we further define a user grouping and power allocation algorithm for NOMA.

2.4 Precoding and User Grouping Algorithm

In this section, we first construct a beamforming vector $\mathbf{w}_n$ for each beam that can effectively reduce or eliminate some interferences. Based on the selected precoding scheme, we further solve the user grouping and power allocation problem, in order to maximize the total system throughput.

2.4.1 Zero-forcing Beamforing

Normally the number of transmit antennas n_T should be larger than or equal to the number of receiver antennas n_R, i.e., $n_T \geq n_R$, so that the transmitter side will have enough degree of freedom to generate a precoding matrix that can effectively eliminate the inter-user interference. In this paper, each MBS has N transmit antennas and can generate N beams. Within each beam, $K(K \geq 2)$ users can be supported by using NOMA. Thus, the total number of receive antennas in this case is $N \times K$, which is larger than N. Existing literatures have observed and addressed this issue. In [18], a coordinated transmit-receive block diagonalization algorithm is put forward. However, the receive antenna set employs a joint precoding matrix, which requires information exchange among different users and consequently adds extra complexity. Here, we consider two zero-forcing precoding methods. The first one aims to minimize the inter-beam interference for CUs while the second one aims to eliminate the interference from MBS to DUs.

First ZF Precoding

In this scheme, we first select one user from each beam and then generate the beamforming matrix based on N selected users. Specifically, users with the largest channel gain in each beam are selected. The channel gain vector for these N selected CUs are denote as $\mathbf{H} = [\mathbf{h}_{u(1,K)}, \mathbf{h}_{u(2,K)} \dots \mathbf{h}_{u(N,K)}]$. The zero-forcing beamforming vector is calculated based on:

$$\mathbf{h}_{u(n,K)}\mathbf{w}_m = 0, \quad \text{if } m \neq n. \tag{2.16}$$

Thus, $\mathbf{w}_m$ should lie in the null space of $\tilde{\mathbf{H}}_n$ [18]. Here, $\tilde{\mathbf{H}}_n$ is defined as

$$\tilde{\mathbf{H}}_n = [\mathbf{h}_{u(1,K)}, \ldots, \mathbf{h}_{u(n-1,K)}, \mathbf{h}_{u(n+1,K)}, \ldots, \mathbf{h}_{u(N,K)}], \tag{2.17}$$

which consists of downlink channel vectors for CUs from all beams except from beam n.

Second ZF Precoding

The first ZF based method helps reduce inter-beam interference $I^U_{u(n,K)} = 0$ in (2.6). Since we aim to maximize the total sum rate in the system, the total throughput from DUs contributes to the total throughput as well. Therefore, the second precoding method helps reduce the interference between CUs and DUs, i.e., $\sum_{n=1}^{N} P_n |\mathbf{h}_p \mathbf{w}_n|^2 = 0$ in (2.10). Hence we should set $\mathbf{h}_p \mathbf{w}_n = 0$, for all n. Or equivalently,

$$\mathbf{w}_n = null(\mathbf{H}_D), \tag{2.18}$$

where $\mathbf{H}_D = [\mathbf{h}_1, \ldots, \mathbf{h}_P]$, and $null(.)$ is the null space or kernel of a matrix.

2.4.2 User Grouping and Optimal Power Allocation

After the beamforming vector is determined, we need to group NOMA users into each beam and further decide power allocation for CUs within each NOMA group. One way is to do an exhaustive search. but the complexity will grow exponentially with N. Inspired by [13] [17], NOMA would prefer to group users with greater channel differences. On the other hand, precoding matrix $\mathbf{W}$ is designed to minimize inter-beam interference or CU to DU interference. When combining NOMA and precoding, NOMA groups users with highly correlated channels so that using the precoding matrix generated by the representative CU in each beam can achieve a small inter-beam or CU-DU interference. Therefore, the criteria for NOMA user grouping is to choose CUs with highly correlated channels but with big channel gain differences in each beam. For simplicity, we set $K = 2$. In each NOMA pair, we denote the user with a weaker channel gain as the first user while the stronger one as

the second user.

First ZF Precoding

Since the bemforming matrix is designed based on the null space of the second users in all N beams, second users will not receive any inter-beam interference. Thus their SINR is

$$\gamma_{u(n,2)} = \frac{\lambda_{u(n,2)} P_n |\mathbf{h}_{u(n,2)}|^2}{I^D_{u(n,2)} + \sigma_n^2}. \tag{2.19}$$

The first users, on the other hand, will receive non-zero inter-beam interference as the precoded signals from other beams will have components projected into the first user signal space. Their SINR is expressed as

$$\gamma_{u(n,1)} = \frac{(1-\lambda_{u(n,2)}) P_n |\mathbf{h}_{u(n,1)} \mathbf{w}_n|^2}{|\mathbf{h}_{u(n,1)} \mathbf{w}_n|^2 \lambda_{u(n,2)} P_n + I^D_{u(n,1)} + I^U_{u(n,1)} + \sigma_n^2}. \tag{2.20}$$

The optimal power allocation factor $\lambda_{u(n,2)}$ is yet to be solved. Based on the optimization problem proposed in the previous section, we form a new problem that aims to maximize the sum capacity in each beam.

$$\max_{\lambda_{u(n,2)}} \sum_{k=1}^{2} f(\mathbb{E}\{\gamma_{u(n,k)}\}) \tag{2.21}$$

subject to

$$0 < \lambda_{u(n,2)} < 1, \tag{2.22}$$

$$f(\mathbb{E}\{\gamma_{u(n,1)}\}) \geq R_0. \tag{2.23}$$

The problem defined above is convex with respect to $\lambda_{u(n,2)}$ and its Karush-Kuhn-Tucker (KKT) conditions are given as follows.

$$\frac{\partial\Big(\sum_{k=1}^{2} f(\mathbb{E}\{\gamma_{u(n,k)}\})\Big)}{\partial \lambda^*_{u(n,2)}} = \mu \frac{\partial\Big(R_0 - f(\mathbb{E}\{\gamma_{u(n,1)}\})\Big)}{\partial \lambda^*_{u(n,2)}}, \tag{2.24}$$

$$R_0 - f(\mathbb{E}\{\gamma_{u(n,1)}\})|_{\lambda^*_{u(n,2)}} \leq 0, \tag{2.25}$$

$$\mu \geq 0, \tag{2.26}$$

$$\mu\Big(R_0 - f(\mathbb{E}\{\gamma_{u(n,1)}\})|_{\lambda^*_{u(n,2)}}\Big) = 0. \tag{2.27}$$

Equation (2.24) is the stationarity condition and μ is KKT multiplier, (2.25) is the primal feasibility, (2.26) is dual feasibility and (2.27) is the complementary slackness.

Solving for (2.24), we can get

$$\lambda^*_{u(n,2)} = \frac{(\mathcal{I}_{D2}+1)\Big((\mathcal{I}_{D1}+1+\Sigma)\mathcal{H}_2 - (1+\mu)\mathcal{H}_1\Big)}{\mathcal{H}_1\mathcal{H}_2\rho(\mu-\mathcal{I}_{D2})}, \tag{2.28}$$

where, $\rho = P_n/\sigma_n^2$ is the transmit SNR, $\mathcal{H}_2 = |\mathbf{h}_{u(n,2)}|^2$, $\mathcal{H}_1 = |\mathbf{h}_{u(n,1)}\mathbf{w}_n|^2$ is the channel gain for user 2 and user 1, $\mathcal{I}_{D1} = I^D_{u(n,1)}/\sigma_n^2$, $\mathcal{I}_{D2} = I^D_{u(n,2)}/\sigma_n^2$ is the interference-to-noise ratio of user 1 and user 2, respectively. $\Sigma = I^U_{u(n,1)}/\sigma_n^2$ is the inter-beam interference-to-noise ratio.

Clearly, $\mu \neq 0$. Otherwise, $\lambda^*_{u(n,2)} < 0$ cannot satisfy (2.22). Therefore, we can solve (2.27) for the optimal $\lambda^*_{u(n,2)}$.

$$\lambda^*_{u(n,2)} = \frac{\rho\mathcal{H}_1 + \mathcal{I}_{D1} + 1 + \Sigma}{2^{R_0}\rho\mathcal{H}_1} - \frac{\mathcal{I}_{D1}+1+\Sigma}{\rho\mathcal{H}_1}. \tag{2.29}$$

$$\lambda^*_{u(n,1)} = 1 - \lambda^*_{u(n,2)}. \tag{2.30}$$

Second ZF Precoding

In the second ZF precoding, the inter-beam interference remains for both first and second users. Their respective SINR are as follows.

$$\gamma_{u(n,2)} = \frac{\lambda_2\rho\mathcal{H}'_2}{\Sigma'_2 + \mathcal{I}'_{D2} + 1}, \tag{2.31}$$

$$\gamma_{u(n,1)} = \frac{(1-\lambda_2)\rho\mathcal{H}'_1}{\lambda_2\rho\mathcal{H}'_1 + \Sigma'_1 + \mathcal{I}'_{D1} + 1}. \tag{2.32}$$

Similarly, λ_2 is the power allocation factor for the user with stronger channel. $\mathcal{H}_1' = |\mathbf{h}_{u(n,1)}\mathbf{w}_{ZF2}|^2$, $\mathcal{H}_2' = |\mathbf{h}_{u(n,2)}\mathbf{w}_{ZF2}|^2$ are the channel gains for user 1 and user 2, respectively. $\Sigma_2' = \sum_{n'=1,n'\neq n}^{N} \rho|\mathbf{h}_{u(n,2)}\mathbf{w}_{ZF2}|^2$ and $\Sigma_1' = \sum_{n'=1,n'\neq n}^{N} \rho|\mathbf{h}_{u(n,1)}\mathbf{w}_{ZF2}|^2$. $\mathcal{I}_{D1}'$ has the same format as $\mathcal{I}_{D1}$ but with different precoding vector, the same to $\mathcal{I}_{D2}'$. We form a similar optimization problem as in (2.21) and detailed derivations are omitted here. The respective optimal power allocation factor for the second and first user is

$$\lambda_2^* = \frac{\rho\mathcal{H}_1' + \mathcal{I}_{D1}' + 1 + \Sigma_1'}{2^{R_0}\rho\mathcal{H}_1'} - \frac{\mathcal{I}_{D1}' + 1 + \Sigma_1'}{\rho\mathcal{H}_1'}, \tag{2.33}$$

$$\lambda_1^* = 1 - \lambda_2^*. \tag{2.34}$$

2.5 Simulation results

In this section, we present the performance results from simulation. The coverage area of MBS is a circle with a radius of 500 m. The number of transmit antennas is $N = 3$. The total numbers of CUs and DUs are $M = [8, 16, 32, 60, 90]$ and $P = 2$ respectively. M varies in order to study the multi-user diversity effect. The distance with each DU pair is fixed at 30 m. The wireless channel consists of pathloss, shadowing and Rayleigh fading with a pathloss exponent 2. P_{MBS} and P_D are set to 30 Watt and 1 Watt, respectively.

For comparison purpose, instead of using NOMA in each beam, we apply a traditional TDMA scheme here to support these 2 users in each beam. Specifically, we allocate an equal number of time slots to 2 TDMA users. The scheme is also referred as "Naive TDMA".

$$R_{TDMA} = \frac{1}{2}\Big(\log(1+\gamma_1) + \log(1+\gamma_2)\Big). \tag{2.35}$$

Fig. 2.2 presents the system capacity of two proposed ZF precoding methods as the number of users grows, the results are scaled over the highest achievable rate. Here we set $R_0 = 0.5$ b/s/Hz. We can see that NOMA outperforms naive TDMA in both precoding schemes when the number of CUs is large. However, when the number is small, limited CUs can be chosen to perform NOMA, thus, the performance gain is not obvious, even worse than

TDMA. We also find that using ZF2 leads to a higher overall system throughput than ZF1. Because with ZF2, DUs experience a much lower interference than with ZF1 so that the throughput elevation from DUs exceeds the throughput degradation from CUs due to inter-beam interference, which results a net gain on overall system throughput. Moreover, as the user number increases, the system benefits more from NOMA+MU-MIMO due to a higher multiuser diversity gain.

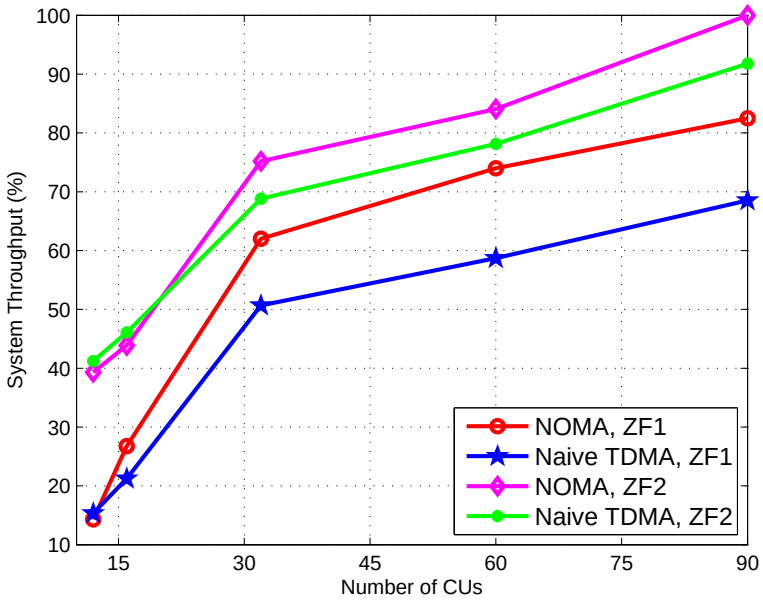

Fig. 2.2: System capacity of two proposed ZF precoding methods vs. TDMA as the number of user grows ($R_0 = 0.5$ b/s/Hz).

DUs normally are considered as a complementary communication method. So we are particularly interested in the performance of CUs. In Fig. 2.3, the throughput of CUs is calculated. NOMA shows a superior spectral efficiency compared with naive TDMA. In this case, ZF1 has a much better performance than ZF2 since ZF1 precoding eliminates inter-beam interference for CUs while ZF2 aims to eliminate interference from CUs to DUs. But if we combine results from both Fig. 2.2 and Fig. 2.3, we can see that the overall throughput is higher with ZF2 since DUs are configured with a very good channel setting so that they contribute to overall throughput significantly.

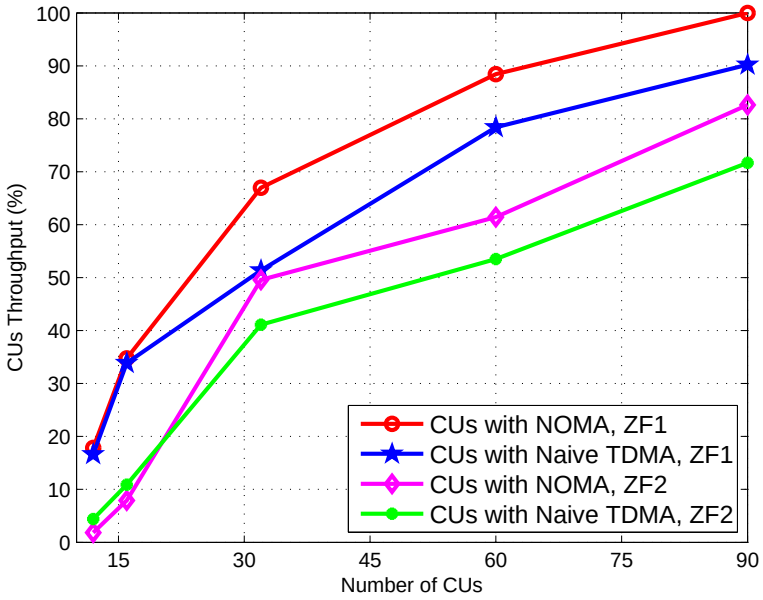

Fig. 2.3: CUs capacity of two proposed ZF precoding methods vs. TDMA as the number of user grows ($R_0 = 0.5$ b/s/Hz).

2.6 Chapter Conclusion

In this chapter, we study the performance of a cellular network that supports NOMA, MU-MIMO and underlay D2D communications. Specifically, we use NOMA and MU-MIMO for the cellular downlink users to improve overall system spectrum efficiency. D2D users are further supported in the underlay mode to exploit the frequency reuse again. Two different precoding mechanisms are defined. We formulate an optimization problem aiming to maximize the system performance and develop a suboptimal approach to solve the problem in two steps. Simulation results show NOMA and MU-MIMO altogether will improve the overall cellular user throughput significantly. When underlay D2D users are added, two different precoding schemes lead to different performance, with one favoring CUEs and one favoring DUEs. But both lead to a net system gain.

CHAPTER 3

Non-orthogonal Multiple Access with SIC Error Propagation in Downlink Wireless MIMO Networks

3.1 Introduction

In the previous chapter, we incorporated NOMA with MIMO in a D2D underlaid system, the advantages of applying NOMA in such a scheme are in two folds: 1) it can support more users simultaneously; 2) overall system performance in terms of total throughput can also be improved. Intuitively, NOMA will create a win-win situation for users with strong and weak channel condition. The reason is that the stronger user is typically bandwidth-limited while the weaker user is interference-limited. In NOMA, signals for both users are set to transmit simultaneously, so the bandwidth-limited user can get more spectrum resources while the interference-limited user can obtain a larger portion of power. This will benefit the whole system in terms of fairness and throughput.

In [14], the concept of NOMA is discussed from the information theoretic perspective, and the conclusion is that NOMA can have a better performance compared with OMA in terms of both system sum rate and user individual rate, especially when the users channel gains are distinct. In [21] and [22], a similar downlink MIMO and NOMA system model is proposed, the authors solve the optimization problem with bisection power search algorithm and use the singular value decomposition (SVD) if the CSI is available at the BS or equally distribute powers among different antennas if CSI is unknown for the precoding design.

These works, however, all assume a perfect subtraction of previous user signals in SIC so that there's no residual interference which will affect the current decoding. This assumption turns out to be a strong one since various factors can actually cause errors, such as deep fading, imperfect decoding and channel estimation errors [23]. In the case of decoding more users' signal, errors from previous will accumulate and greatly affect the next stage (we refer

this as error propagation). In this chapter, we take error propagation into consideration, a concept that already exists in CDMA systems. Similar papers can be found in [24] and [25]. In fact, CDMA shares some common features with NOMA. Both of them exploit the multiuser interference to achieve a higher performance rather than simply avoid it. Performance gain also largely depends on some assumptions like perfect channel estimation and power allocation, and violations can cause serious performance degradation. In this paper, we propose a general error propagation model in a downlink MIMO NOMA system, where decoding errors are modelled as residual interference. An optimization problem is formulated to maximize the total data rate of two users.

3.2 System Model

We consider a downlink wireless communication which jointly supports NOMA and MU-MIMO. In the system, a BS with power P_{BS} is equipped with M antennas. Two UEs are randomly deployed in this area, each has N antennas.

Due to the usage of NOMA, two UEs can receive signals from the BS simultaneously. Besides, the BS is assumed to have an accurate CSI of UEs based on training sequences and feedback mechanism. We denote $\mathbf{H}_k$ and $\mathbf{H}_n$(both with dimension $\mathbb{C}^{N\times M}$) as the channel gain matrix of UE k and UE n, respectively. h_{ij} is the element from ith row and jth column in the matrix and it is modeled as the product of large-scale path loss and fading, i.e, $h_{ij} = l_{ij}^{-\alpha} h_0$, where l_{ij} is the distance between UE and BS, α is the path loss exponent and h_0 is the Gaussian random variable with distribution $h_0 \sim \mathcal{CN}(0, 1)$.

The transmitted signal from the BS is:

$$\mathbf{x}_{BS} = \mathbf{W}_n \mathbf{x}_n + \mathbf{W}_k \mathbf{x}_k, \tag{3.1}$$

where $\mathbf{W}_n$ and $\mathbf{W}_k$ are precoding matrices with dimension $\mathbb{C}^{M\times N}$, $\mathbf{x}_n$ and $\mathbf{x}_k \in \mathbb{C}^{N\times 1}$ are messages for UE n and UE k, respectively. $\mathbb{E}(\mathbf{x}_n \mathbf{x}_n^H) = \mathbb{E}(\mathbf{x}_n \mathbf{x}_n^H) = \mathbf{I}_N$, $\mathbb{E}(.)$ is the expectation function and $\mathbf{I}_N$ is a $N \times N$ identity matrix.

The received signal at UE n is

$$\mathbf{y}_n = \mathbf{H}_n \mathbf{x}_{BS} + \mathbf{n}_n. \tag{3.2}$$

Similarly, UE k will receive,

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x}_{BS} + \mathbf{n}_k, \tag{3.3}$$

where $\mathbf{n}_i, i = \{n, k\}$ is the i.i.d additive gaussian noise which follows $\mathcal{CN}(0, \sigma^2 \mathbf{I}_N)$.

3.3 SIC and Problem Formulation

The key idea of SIC can be summarized as a process of decoding, reconstruction and subtraction (DRS). Upon the reception of the composite signal, DRS will start with the strongest user signal and treat others as interference. After the successful decoding, the data will re-encode based on user channel estimation and constellation. The reconstructed signal should be fairly close to the received signal if everything is perfect. Then, the user will subtract this signal from the aggregated signal so that the next DRS will see less interference if the intended message is not decoded [10]. However, DRS can be affected by error propagation, we will show this concept below.

3.3.1 SIC with Error Propagation

Sequential decoding can be affected by error propagation. Consider a simpler system with one BS and two UEs, in which UE 1 and UE 2 form a NOMA pair. The power of the BS is P and the channel gains for UE 1 and UE 2 are h_1 and h_2, respectively. Without loss of generality, let $h_1 > h_2$. The transmitted signal can be expressed as,

$$x_t = \sqrt{\theta_1 P} s_1 + \sqrt{\theta_2 P} s_2, \tag{3.4}$$

where θ_1 and θ_2 are power allocation factors, $\theta_1 < \theta_2$ for QoS consideration and $\theta_1 + \theta_2 = 1$. s_1 and s_2 are normalized signals.

At the receiver side, UE 1 will get $y_1 = h_1 x_t + n_1 = h_1(\sqrt{\theta_1 P} s_1 + \sqrt{\theta_2 P} s_2) + n_1$. Clearly

the signal for UE 2 has a larger power than that for UE 1, thus, at the first stage, UE 1 will decode UE 2's signal. Let $R_{1,2}$ denote the achievable data rate for UE 1 to detect UE 2's message, it can be expressed as,

$$R_{1,2} = \log_2 \left(1 + \frac{\theta_2 P|h_1|^2}{\theta_1 P|h_1|^2 + n_1^2}\right). \tag{3.5}$$

UE 1 then reconstructs this message according to a prior known constellation and channel gain. After that UE 1 will subtract UE 2's signal and decode its own, and the data rate is given by,

$$R_1 = \log_2 \left(1 + \frac{\theta_1 P|h_1|^2}{n_1^2}\right). \tag{3.6}$$

The received signal for UE 2 is $y_2 = h_2 x_t + n_2 = h_2(\sqrt{\theta_1 P}s_1 + \sqrt{\theta_2 P}s_2) + n_2$, since the desired signal has a larger power, so it can be detected directly. The achievable data rate for UE 2 is simply

$$R_2 = \log_2 \left(1 + \frac{\theta_2 P|h_2|^2}{\theta_1 P|h_2|^2 + n_2^2}\right). \tag{3.7}$$

The above procedure, however, depends on the perfect DRS of UE2's signal at UE 1, which is a strong assumption, since various factors such as deep fading can affect the signal detection and decoding. Assuming at UE 1 side, the DRS procedure is not perfect, thus there will be residual signal power at stage 2 when decoding its own message. As a result, the data rate for UE 1 becomes,

$$R_1^{'} = \log_2 \left(1 + \frac{\theta_1 P|h_1|^2}{\beta\theta_2 P|h_1|^2 + n_1^2}\right), \tag{3.8}$$

where β is the error propagation factor, which is inversely proportional to the SINR of (3.5), i.e., $\beta \propto \frac{\theta_1 P|h_1|^2 + n_1^2}{\theta_2 P|h_1|^2}$ and $0 \leq \beta \leq 1$. $\beta = 0$ represents the perfect decoding, which is the same as (3.6). While $\beta = 1$ is the worst case that the DRS of UE2 is totally unsuccessful and UE1 has to treat its entire signal as interference. (In this case, it has the same result as without SIC.)

In our system model, if we assume $\mathbf{H}_n\mathbf{H}_n^H \succ \mathbf{H}_k\mathbf{H}_k^H$, here $\succ$ means if $\mathbf{A} \succ \mathbf{B}$, then

$(\mathbf{A}-\mathbf{B})$ is a positive definite matrix. This assumption implies UE n has a better channel condition and hence can decode UE k's message. Thus, at UE n, we have

$$R_{n,k} = \log_2 \det\left(\mathbf{I} + (\sigma^2\mathbf{I} + \mathbf{H}_n\mathbf{W}_n\mathbf{W}_n^H\mathbf{H}_n^H)^{-1}\mathbf{H}_n\mathbf{W}_k\mathbf{W}_k^H\mathbf{H}_n^H\right), \tag{3.9}$$

which is the maximum achievable rate for UE k at UE n. Considering the error propagation, the data rate for UE n's own message would be,

$$R_n{=}\log_2 \det\left(\mathbf{I} + (\sigma^2\mathbf{I} + \beta\mathbf{H}_n\mathbf{W}_k\mathbf{W}_k^H\mathbf{H}_n^H)^{-1}\mathbf{H}_n\mathbf{W}_n\mathbf{W}_n^H\mathbf{H}_n^H\right). \tag{3.10}$$

The error propagation factor β is assumed be a fixed value.

While at UE k, the desired signal can be decoded directly.

$$R_{k,k} = \log_2 \det\left(\mathbf{I} + (\sigma^2\mathbf{I} + \mathbf{H}_k\mathbf{W}_n\mathbf{W}_n^H\mathbf{H}_k^H)^{-1}\mathbf{H}_k\mathbf{W}_k\mathbf{W}_k^H\mathbf{H}_k^H\right). \tag{3.11}$$

In order for UE k to have a fairly small bit error rate (BER), the maximum allowable data rate for UE k is,

$$R_k = \min\{R_{n,k}, R_{k,k}\}. \tag{3.12}$$

Here we normalize the bandwidth at the BS to 1.

Next, we show that $R_k = R_{k,k}$, the proof follows appendix A in [21] and can be briefly summarized as follows.

Proof. Since $\mathbf{H}_n\mathbf{H}_n^H \succ \mathbf{H}_k\mathbf{H}_k^H$, we can write $\mathbf{H}_n = \mathbf{M}\mathbf{H}_k$, where $\mathbf{M}$ is a $N \times N$ matrix and $\mathbf{M}\mathbf{M}^H \succ \mathbf{I}_N$.

Due to the property of determinant operation, we can rewrite $R_{n,k}$ as

$$R_{n,k} = \log_2 \det\left(\mathbf{I} + \mathbf{W}_k^H\mathbf{H}_n^H(\sigma^2\mathbf{I} + \mathbf{H}_n\mathbf{W}_n\mathbf{w}_n^H\mathbf{H}_n^H)^{-1}\mathbf{H}_n\mathbf{W}_k\right) \tag{3.13}$$

Define $\mathbf{Q}_{n,k} = \mathbf{W}_k^H \mathbf{H}_n^H (\sigma^2 \mathbf{I} + \mathbf{H}_n \mathbf{W}_n \mathbf{W}_n^H \mathbf{H}_n^H)^{-1} \mathbf{H}_n \mathbf{W}_k$ and
$\mathbf{Q}_{k,k} = \mathbf{W}_k^H \mathbf{H}_k^H (\sigma^2 \mathbf{I} + \mathbf{H}_k \mathbf{W}_n \mathbf{W}_n^H \mathbf{H}_k^H)^{-1} \mathbf{H}_k \mathbf{W}_k$, then we substitute $\mathbf{H}_n = \mathbf{M}\mathbf{H}_k$ in $\mathbf{Q}_{n,k}$.

$$\begin{aligned}
\mathbf{Q}_{n,k} &= \mathbf{W}_k^H \mathbf{H}_n^H (\sigma^2 \mathbf{I} + \mathbf{M}\mathbf{H}_k \mathbf{W}_n \mathbf{W}_n^H \mathbf{H}_k^H \mathbf{M}^H)^{-1} \mathbf{H}_n \mathbf{W}_k \\
&= \mathbf{W}_k^H \mathbf{H}_k^H (\sigma^2 (\mathbf{M}^H \mathbf{M})^{-1} + \mathbf{H}_k \mathbf{W}_n \mathbf{W}_n^H \mathbf{H}_k^H)^{-1} \mathbf{H}_k \mathbf{W}_k \\
&\succ \mathbf{W}_k^H \mathbf{H}_k^H (\sigma^2 \mathbf{I} + \mathbf{H}_k \mathbf{W}_n \mathbf{w}_n^H \mathbf{H}_k^H)^{-1} \mathbf{H}_k \mathbf{W}_k \\
&= \mathbf{Q}_{k,k} \qquad (3.14)
\end{aligned}$$

Thus, $\log_2 \det(\mathbf{I} + \mathbf{Q}_{n,k}) > \log_2 \det(\mathbf{I} + \mathbf{Q}_{k,k})$, which means $R_{n,k} > R_{k,k}$, so $R_k = R_{k,k}$. □

3.3.2 Problem Formation

In this chapter, we intend to maximize the system throughput by applying NOMA and MU-MIMO. The problem can be formed as following.

$$\max_{\mathbf{W}_n, \mathbf{W}_k} (R_n + R_k) \qquad (3.15)$$

subject to

$$\mathrm{tr}(\mathbf{W}_n \mathbf{W}_n^H + \mathbf{W}_k \mathbf{W}_k^H) \leq P_{BS}, \qquad (3.16)$$

$$R_k \geq R_0. \qquad (3.17)$$

(3.16) is the constraint for maximum allowed power from the BS. (3.17) sets a minimum data rate for the weaker UE. R_n and R_k can be calculated based on (3.10) and (3.11), respectively. One note here is due to the error propagation, the stronger UE may suffer severe residual interference from the weaker one, thus its data rate may be lower. However, we do not consider this situation in the paper, the lower data rate limit is only for the weaker UE.

When a resource block (RB) is available, the BS needs to determine the following: 1) How to properly design the precoding matrix; 2) How to allocate the power to each UE.

The above optimization problem is hard to solve, the reason is that it imposes the

error propagation, which makes the utility function (3.15) hard to track. Besides, when calculating R_i, we also need to determine the precoding matrix $\mathbf{W}_i$, for $i = n, k$. In the next section, we propose an unified precoding matrix formation algorithm, then we focus on the power allocation with residual interference.

3.4 Precoding and Power Allocation

3.4.1 Precoding Design

Let $\mathrm{tr}(\mathbf{W}_n\mathbf{W}_n^H) = P_n$ and $\mathrm{tr}(\mathbf{W}_k\mathbf{W}_k^H) = P_k$. The optimization problem can be revised as:

$$\max_{\mathbf{W}_n, \mathbf{W}_k} (R_n + R_k) \tag{3.18}$$

subject to

$$\mathrm{tr}(\mathbf{W}_n\mathbf{W}_n^H) = P_n, \tag{3.19}$$

$$\mathrm{tr}(\mathbf{W}_k\mathbf{W}_k^H) \leq P_{BS} - P_n, \tag{3.20}$$

$$R_k \geq R_0. \tag{3.21}$$

The introduced error propagation model increases the complexity of the optimization problem. Basically it is a MIMO broadcast channel (BC) in the downlink. So it can be converted to multiple access channel (MAC) in the uplink using BC-MAC duality. But it requires extensive matrix calculation and is not easy to understand. Here in this paper, we introduce the equivalent channel and its respective precoding solution.

From (3.10), we denote $\mathbf{H}n_{eq}$ as the equivalent channel of UE n and it can be expressed as $(\sigma^2\mathbf{I}+\beta\mathbf{H}_n\mathbf{W}_k\mathbf{W}_k^H\mathbf{H}_n^H)^{-1/2}\mathbf{H}_n$. We then rewrite (3.10) in terms of the equivalent channel $\mathbf{H}n_{eq}$.

$$R_n = \log_2 \det(\mathbf{I} + \mathbf{H}n_{eq}\mathbf{W}_n\mathbf{W}_n^H\mathbf{H}n_{eq}^H). \tag{3.22}$$

Similarly, (3.11) can be expressed as,

$$R_k = \log_2 \det(\mathbf{I} + \mathbf{H}k_{eq}\mathbf{W}_k\mathbf{W}_k^H\mathbf{H}k_{eq}^H), \tag{3.23}$$

where $\mathbf{H}k_{eq} = (\sigma^2\mathbf{I} + \mathbf{H}_k\mathbf{W}_n\mathbf{W}_n^H\mathbf{H}_k^H)^{-1/2}\mathbf{H}_k$.

Thus, we can treat the problem as two point-to-point MIMO UEs with a total power constrain, which is already well known in the literature [26]. However, due to the imposed minimum data rate requirement for UE k, It is not necessarily the optimal solution. The suboptimal precoding can be formed as follows. First, take the SVD of the equivalent channel,

$$\mathbf{U}_n\mathbf{E}_n\mathbf{U}_n^H = \mathbf{H}n_{eq}^H\mathbf{H}n_{eq}, \tag{3.24}$$

where $\mathbf{U}_n$ is the unitary matrix and its columns are a set of orthonormal eigenvectors of $\mathbf{H}n_{eq}^H\mathbf{H}n_{eq}$, $\mathbf{E}_n$ is a diagonal matrix. Therefore, the precoding matrix can be formed as,

$$\mathbf{W}_n\mathbf{W}_n^H = \mathbf{U}_n\widetilde{\mathbf{E}_n}\mathbf{U}_n^H, \tag{3.25}$$

where $\widetilde{\mathbf{E}_n}$ is calculated from water-filling process with respect to the elements in the diagonal matrix $\mathbf{E}_n$, i.e., $\widetilde{\mathbf{E}_n} = \lceil \lambda_n\mathbf{I} - (\mathbf{E}_n)^{-1} \rceil^+$, here λ_n is a parameter to ensure the power constraint $\mathrm{tr}(\mathbf{W}_n\mathbf{W}_n^H) = P_n$, and $\lceil a \rceil^+ = \max(a, 0)$.

The precoding matrix for UE k can be formed in the same way. However, two problems remain here: 1) The calculation of $\mathbf{W}_n$ involves $\mathbf{W}_k$ and vice versa; 2) Power P_n and P_k are unknown. Next, we propose an iterative way to solve for precoding generation under the assumption that each UE's power is known as *a prior*.

We start with $\mathbf{W}_k\mathbf{W}_k^H = \frac{P_k}{M}\mathbf{I}_M$, and calculate $\mathbf{H}n_{eq}$, $\mathbf{W}_n\mathbf{W}_n^H$ and $\mathbf{H}k_{eq}$ sequentially, then we update $\mathbf{W}_k\mathbf{W}_k^H$ according to the new $\mathbf{H}k_{eq}$. The process will continue until reaches the maximum iteration number. To make further clarification, the algorithm for precoding design is summarized in **Algorithm** 1.

A note here is that the covariance matrix $\mathbf{W}_k\mathbf{W}_k^H$ and $\mathbf{W}_n\mathbf{W}_n^H$ are actually characterized the data rate, not $\mathbf{W}_n$ or $\mathbf{W}_k$ individually. And an easy way to find $\mathbf{W}_n$ and $\mathbf{W}_k$

Algorithm 1 Iterative Precoding Design

1: **Initialization:** Given power P_n and P_k, maximum iteration number $MAXITER$.
2: $\mathbf{W}_k\mathbf{W}_k^H = \frac{P_k}{M}\mathbf{I}_M$.
3: **for** $i = 1$ to $MAXITER$ **do**
4: Calculate $\mathbf{H}n_{eq}$ based on $\mathbf{W}_k\mathbf{w}_k^H$
5: Solve for $\mathbf{W}_n\mathbf{W}_n^H$ from the SVD of $\mathbf{H}n_{eq}^H\mathbf{H}n_{eq}$.
6: Calculate $\mathbf{H}k_{eq}$ based on $\mathbf{W}_n\mathbf{W}_n^H$
7: Update $\mathbf{W}_k\mathbf{W}_k^H$ from the SVD of $\mathbf{H}k_{eq}^H\mathbf{H}k_{eq}$.
8: **end for**
9: Output R_n, R_k, $\mathbf{W}_k\mathbf{W}_k^H$ and $\mathbf{W}_n\mathbf{W}_n^H$.

is,

$$\mathbf{W}_n = \mathbf{U}_n\widetilde{\mathbf{E}_n}^{\frac{1}{2}}, \mathbf{W}_k = \mathbf{U}_k\widetilde{\mathbf{E}_k}^{\frac{1}{2}}, \tag{3.26}$$

which is rather straightforward.

3.4.2 Case Studies for Power Allocation

In this subsection, two case studies are investigated.

Case I

The error propagation factor β is a small value. In this special case, we can omit the impact of imperfect DRS process and R_n becomes,

$$R_n = \log_2 \det\left(\mathbf{I} + (\sigma^2\mathbf{I})^{-1}\mathbf{H}_n\mathbf{W}_n\mathbf{W}_n^H\mathbf{H}_n^H\right). \tag{3.27}$$

Since the sum rate $(R_n + R_k)$ is an monotone increasing function of P_n, as shown in [21], so we only need to find the minimum power for the weak user k to meet the data rate requirement, then allocate the rest power to UE n. In this case, we can get the optimal power by using bisection search algorithm [21] [22].

Case II

β is large. In this case, we may discard the ambient (thermal) noise and R_n is only affected by the residual interference from UE k.

$$R_n = \log_2 \det \left(\mathbf{I} + (\beta \mathbf{H}_n \mathbf{W}_k \mathbf{W}_k^H \mathbf{H}_n^H)^{-1} \mathbf{H}_n \mathbf{W}_n \mathbf{W}_n^H \mathbf{H}_n^H \right). \tag{3.28}$$

This can happen when the received SINR for UE k is relatively small, causing a higher error probability. The sum rate in this case in neither an increasing or decreasing function of P_n, and hence is difficult to track. As we will see later in the simulation section, sum rate is affected by the choice of β. As a preliminary research, we present some results on how the power allocation will affect the sum rate.

3.5 Simulation Analysis

In this section, we present our simulation results. The total power of the BS is 2 Watts. The number of BS and UE antennas are both equal to 2. The average channel gain for UE n and UE k are 0 and 5 dB, respectively. As for the small β, we choose $\beta = 0.05$, while the large β equals to 0.65. $\sigma = 0.5$ in our system. The minimum data rate for UE k is 1 bits/s/Hz. For comparison purposes, we also list the results with precoding as $\mathbf{W}_k \mathbf{W}_k^H = \frac{P_k}{M} \mathbf{I}_M$ and $\mathbf{W}_n \mathbf{W}_n^H = \frac{P_n}{M} \mathbf{I}_M$. $MAXITER = 5$ as our iterative precoding algorithm converges very fast. All the results come from 10,000 independent Monte Carlo experiments to ensure the confidence level.

Fig. 3.1 shows the rates of UE n and UE k as P_n changes, respectively. β is set to be 0.05 for error propagation in this case. We can see that the rate of UE n increases when P_n increases while the rate of UE k decreases when P_n increases. It is obvious that the rate of a UE increase when its assigned power increases since the SINR increases. We can also see the rate of UE n increases faster than the rate of UE k when their power increase individually. Since UE n is a user with better channel condition, increasing power slightly can increase the rate a lot. Since β is small, the residual interference from UE k does not affect the performance of UE n too much. UE rates are also shown for identity matrix

precoding method. The performance of the identity matrix precoding method has the similar trend to the performance of the proposed precoding design, but the identity matrix precoding method does not perform as well as the proposed precoding design. Another note is the gap between two precoding matrices is small with UE n, this is because as the SINR increases, the water-filling algorithm has a similar performance compared with equal power distribution.

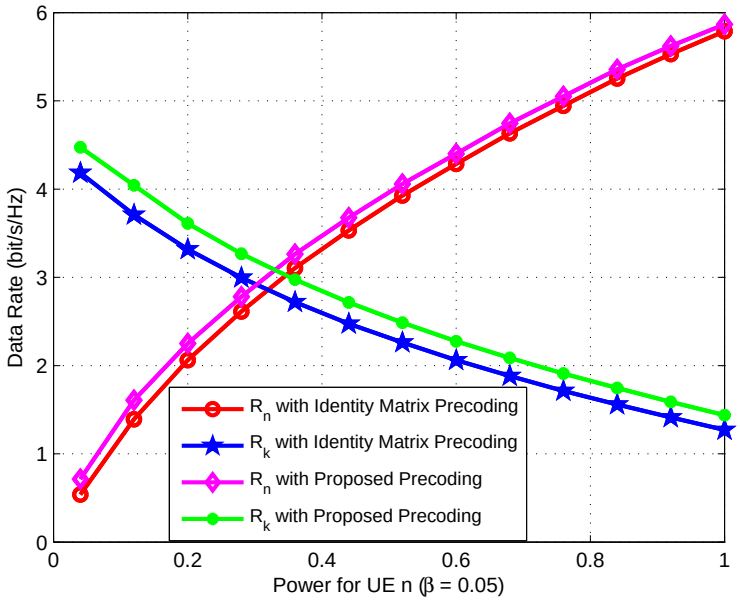

Fig. 3.1: UE rate with different precoding matrix as P_n increases. ($\beta = 0.05$)

Fig. 3.2 shows the sum rate of UE n and UE k as P_n changes when $\beta = 0.05$. We can see that the sum rate increases while P_n increases. From Fig. 3.1, it has been shown that the rate of UE n increases faster than the decreasing speed of rate of UE k when P_n increases. Therefore, the sum rate of UE n and UE k increases while P_n increases. Fig. 3.2 also shows the proposed precoding design performs better than identity matrix precoding method with respect to sum rate.

Fig. 3.3 shows the rates of UE n and UE k as P_n changes when β is set to be 0.65 for error propagation. We can still see that the rate of UE n increases when P_n increases

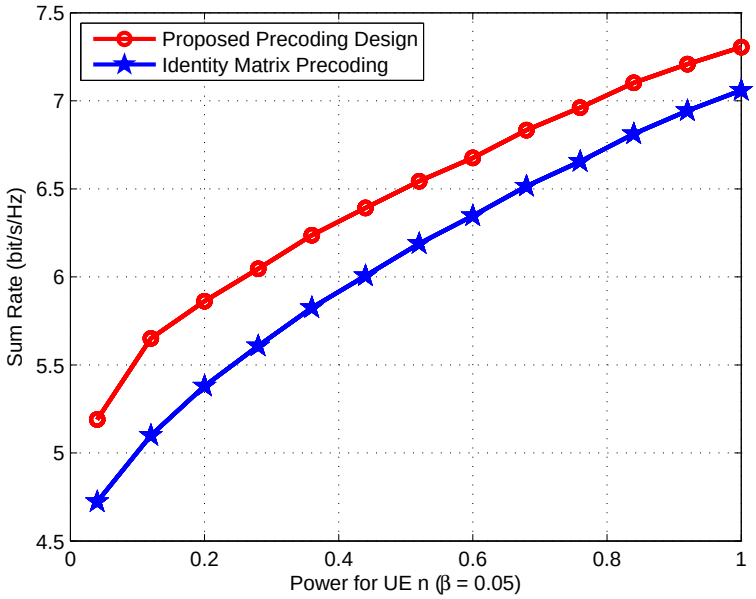

Fig. 3.2: Sum rate with different precoding matrix as P_n increases. ($\beta = 0.05$)

while the rate of UE k decreases when P_n increases. However, the rate of UE n increases faster than the decreasing speed of rate of UE k when P_n increases. The rate of UE n increases slower than the decreasing speed of rate of UE k when $P_n > 0.76$. The reason is that when $P_n < 0.76$, the interference from UE n is smaller than the noise, therefore the SINR decreases slowly. However, when $P_n > 0.76$, the interference becomes dominant and causes the SINR to decrease rapidly. Compare with Fig. 3.1, the rate of UE n increases slower than that in Fig. 3.1. Since a bigger β is used in this case, residual interference from UE k has a bigger effect to UE n. Therefore, the rate of UE n increases slower because of stronger residual interference from UE k. For the identity matrix precoding method, the rate of UE n increases slower than the rate of UE k when their power increases individually because UE n has a strong residual interference from UE k. We can see that the proposed precoding design performs much better than the identity matrix precoding method because it is designed to optimize the UE sum rate.

Fig. 3.4 shows the sum rate of UE n and UE k as P_n changes when $\beta = 0.65$. We can see that the proposed precoding design performs much better than the identity matrix

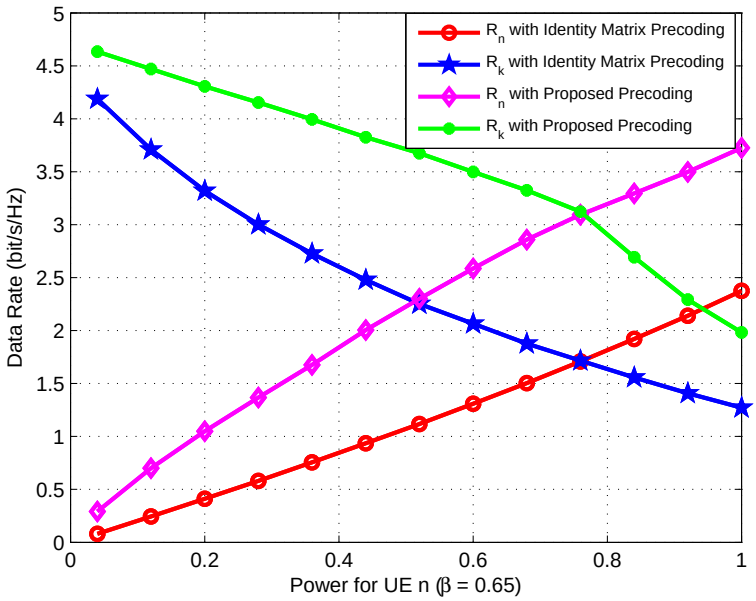

Fig. 3.3: UE rate with different precoding matrix as P_n increases. ($\beta = 0.65$)

precoding method because it is designed to optimize the UE sum rate.

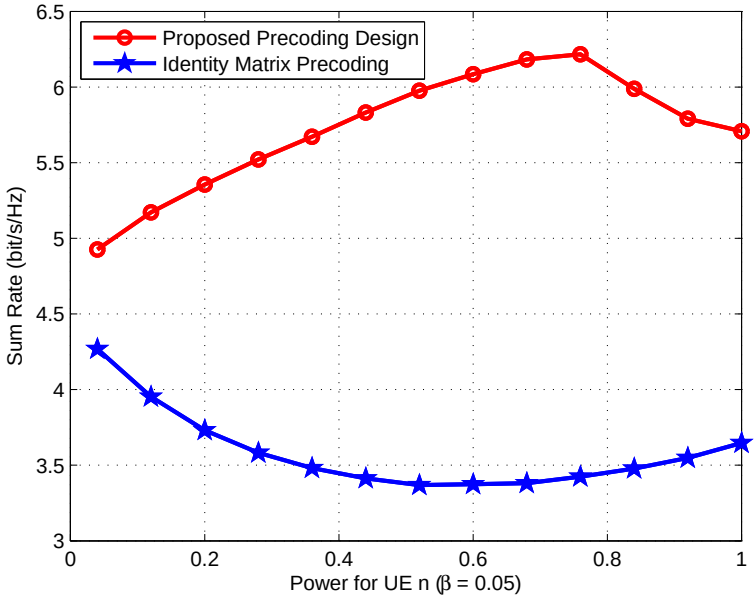

Fig. 3.4: Sum rate with different precoding matrix as P_n increases. ($\beta = 0.65$)

3.6 Chapter Conclusion

In this chapter, we consider a downlink wireless network which jointly incorporates NOMA and MIMO. A sum rate optimization problem is formulated with error propagation in SIC. In order to solve the problem, we introduce the concept of equivalent channel and propose a sequential solution which solves for precoding matrix by applying an iterative algorithm first. Then we investigate the impact of power allocation by two cases: small error propagation factor and large error propagation factor. Simulations are performed to verify the superiority of proposed precoding design and our analyses on power allocation with residual interference.

CHAPTER 4

NOMA in Relay and IoT Networks

4.1 Outage Probability Study in a NOMA Relay System

4.1.1 Introduction

In previous chapters, we have shown NOMA in D2D underlaid MIMO networks and the SIC error propagation. Existing works also evaluated NOMA's performance under outage analysis. In [14], the authors analyzed the performance of NOMA theoretically and they concluded that the disparity, either from user channels or intentionally created by allocating different power factor, can further be beneficial to the system performance. A similar conclusion was drawn from CR-NOMA in [13]. Outage probability is a metric widely used in performance evaluation. It is shown in [8] that the outage performance of NOMA is superior to the traditional OMA in a group of randomly deployed users.

This chapter develops a precoding and power allocation strategy to further enhance the system performance in terms of sum rate. Similarly, both [21] and [22] apply NOMA into MIMO scheme. The algorithms in their papers can be applied with or without CSI. In [9] and [15] system-level performance of using NOMA in LTE and heterogeneous networks is evaluated and the results show promising improvements over existing radio access technologies (RATs). In [29], random beamforming together with intra-beam superposition coding and SIC with BS cooperation is investigated. Relay cooperative communication has been studied in the following papers. [30] uses a single-antenna amplified-and-forward (AF) relay to help the transmission between multi-antenna BS and users. [31] uses relay to help the transmission to a poor-channel user. [32] investigates the system performance under a selection of multiple relays.

As NOMA uses SIC at the receiver to decode multiple user information, the performance of SIC can greatly impact NOMA. Most existing papers assume perfect SIC in NOMA study with a few like [23] considering SIC error due to imperfect channel knowledge. One of our earlier papers [33] investigates the sum rate performance in a MIMO+NOMA system and it considers error propagation in the SIC process. The idea was inspired by the decoding process in CDMA systems [24]. It assumes there is a residual power from previously decoded signals and this residual power can arise due to channel estimation error, imperfect constellation mapping, or channel fading. SIC error propagation causes a chain effect and it affects the last decoded user most.

In this chapter, two NOMA relay schemes are presented and evaluated, namely NOMA cooperative scheme and NOMA TDMA scheme. In NOMA cooperative scheme, the completion of one round information transmission consists of two time slots. In the first slot the BS uses NOMA to send the superimposed signal to two relays. Upon receiving the signal, these two relays will decode the signals by using SIC and then form a cooperative communication pair to send the precoded signals to the respective recipients in the second time slot. DPC is used as precoding at relays to eliminate the inter-user interference in the second time slot. As a comparison, NOMA TDMA scheme uses three times slots to complete one round information transmission. The first time slot does the same as in scheme one. After relays decode the message, the first relay sends one signal to user one in the second slot and the second relay send another signal to user 2 in the following slot. Analytical models on outage performance are derived for both schemes in this paper.

4.1.2 System Model

The study considers a downlink wireless communication system that consists of one access point (AP), and a number of UEs. Each UE can either function as a relay when needed or as a regular UE. The transmit powers of AP and UE are P_s and P_r, respectively. By using NOMA, the AP can communicate with two UEs simultaneously. In the case that the channels between AP and these two UEs are poor, two other UEs are selected as relays for multi-hop cooperative transmission. Relays operate in a half-duplex

decode-and-forward (DF) mode. The AP and UEs in the system are equipped with a single antenna. For notational simplicity, we denote AP, relay 1, relay 2, UE 1 and UE 2 by using subscripts $b, r1, r2, u1$ and $u2$ in the equations, respectively. Furthermore, it is assumed that channels between the AP and two relays are two independent random variables (RVs) following a complex Gaussian distribution with zero mean but different variance, i.e., $h_{b,r1} \sim \mathcal{CN}(0, \sigma^2_{b,r1})$, $h_{b,r2} \sim \mathcal{CN}(0, \sigma^2_{b,r2})$. Without loss of generality, we assume $|h_{b,r1}|^2 > |h_{b,r2}|^2$ and thus $\alpha_s < \beta_s$ is satisfied to provide sufficient decoding capability for NOMA weaker user. On the other hand, channels between relays and UEs can be modeled as independent complex Gaussian RVs with zero mean and unit variance, i.e. $h_{i,j} \sim \mathcal{CN}(0, 1). i = \{r1, r2\}, j = \{u1, u2\}$.

NOMA Cooperative Scheme

Each round of NOMA cooperative transmission consists of two time slots. In the first time slot, the AP transmits a composite signal $x_s = \sqrt{\alpha_s P_s} x_1 + \sqrt{\beta_s P_s} x_2$ according to the NOMA principle, where x_1 and x_2 are signals intended for user 1 and user 2 respectively; α_s and β_s are the corresponding power allocation factors and satisfy $\alpha_s + \beta_s = 1$. The received signals at two relays are respectively expressed as

$$y_{b,r1} = h_{b,r1}(\sqrt{\alpha_s P_s} x_1 + \sqrt{\beta_s P_s} x_2) + n_{b,r1} \tag{4.1}$$

and

$$y_{b,r2} = h_{b,r2}(\sqrt{\alpha_s P_s} x_1 + \sqrt{\beta_s P_s} x_2) + n_{b,r2}. \tag{4.2}$$

$n_{b,r1}$ and $n_{b,r2}$ are additive white Gaussian noise (AWGN) and follow $n_{b,i} \sim \mathcal{CN}(0, N_0), i = \{r1, r2\}$. Both relays use SIC to decode the received signals. We first present the analysis by assuming perfect SIC and the results with imperfect SIC will be presented later. For

relay 1, x_2 will be decoded first by treating x_1 as interference and the achievable signal-to-noise-plus-interference ratio (SINR) for x_2 is

$$\gamma_{r1,x2} = \frac{\beta_s P_s |h_{b,r1}|^2}{\alpha_s P_s |h_{b,r1}|^2 + N_0}. \tag{4.3}$$

Relay 1 then subtracts x_2 from the composite signal and decodes x_1 with only AWGN. Thus, the achievable SINR becomes

$$\gamma_{r1,x1} = \frac{\alpha_s P_s |h_{b,r1}|^2}{N_0}. \tag{4.4}$$

Similarly, at relay 2, the SINR for x_2 and x_1 can be expressed as

$$\gamma_{r2,x2} = \frac{\beta_s P_s |h_{b,r2}|^2}{\alpha_s P_s |h_{b,r2}|^2 + N_0}, \tag{4.5}$$

and

$$\gamma_{r2,x1} = \frac{\alpha_s P_s |h_{b,r2}|^2}{N_0}, \tag{4.6}$$

respectively.

In the second time slot, relay 1 transmits x_1 to user 1 while relay 2 transmits x_2 to user 2 by using precoded cooperative transmission. The received signals at user 1 and user 2 are expressed as

$$\begin{aligned} y_{u1} &= h_{r1,u1}\hat{x}_1 + h_{r2,u1}\hat{x}_2 + n_{u1}, \\ y_{u2} &= h_{r1,u2}\hat{x}_1 + h_{r2,u2}\hat{x}_2 + n_{u2}, \end{aligned} \tag{4.7}$$

where AWGN $n_i \sim \mathcal{CN}(0, N_0), i = \{u1, u2\}$. If we re-write the above equation in the matrix format, we can get $\mathbf{y} = \mathbf{H}\hat{\mathbf{x}} + \mathbf{n}$ and $\mathbf{y} = [y_{u1} \quad y_{u2}]^T$. $\hat{\mathbf{x}} = [\hat{x}_1 \quad \hat{x}_2]^T$ is the precoded transmitted signal vector. The precoding mechanism will be discussed later.

$\mathbf{n} = [n_{u1} \quad n_{u2}]^T$ and

$$\mathbf{H} = \begin{bmatrix} h_{r1,u1} & h_{r2,u1} \\ h_{r1,u2} & h_{r2,u2} \end{bmatrix}. \tag{4.8}$$

To further minimize inter-user interference, DPC is applied at relays as the precoding scheme. Assume $\mathbf{H}$ is a full-rank matrix and it can be decomposed as $\mathbf{H} = \mathbf{LQ}$, where $\mathbf{L}$ is a 2×2 lower triangular matrix and $\mathbf{Q}$ is a semi-orthogonal matrix, $\mathbf{QQ}^H = \mathbf{I}_2$. Thus, let $\mathbf{W} = \mathbf{Q}^H\mathbf{G}$ and $\mathbf{G}$ is given as

$$\mathbf{G} = \begin{bmatrix} 1 & 0 \\ -\frac{l_{2,1}}{l_{2,2}} & 1 \end{bmatrix}, \tag{4.9}$$

where $l_{i,j}$ is the (i,j)-th entry of matrix $\mathbf{L}$.

The received signals at two users can be expressed as

$$\begin{aligned} \mathbf{y} &= \mathbf{H}\hat{\mathbf{x}} + \mathbf{n} = \mathbf{HWx} + \mathbf{n} \\ &= \begin{bmatrix} l_{1,1} & 0 \\ 0 & l_{2,2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} n_{u1} \\ n_{u2} \end{bmatrix}. \end{aligned} \tag{4.10}$$

Therefore, SINRs for user 1 and user 2 can be written as

$$\gamma_{u1} = \frac{|l_{1,1}|^2 P_r}{N_0}, \gamma_{u2} = \frac{|l_{2,2}|^2 P_r}{N_0}. \tag{4.11}$$

An illustration of NOMA cooperative scheme is shown in Fig. 4.1.

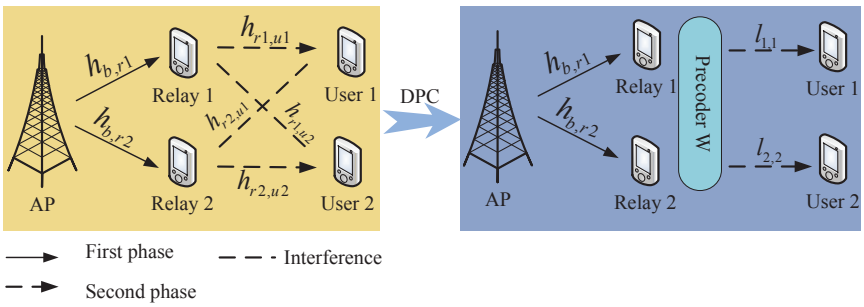

Fig. 4.1: NOMA cooperative scheme

For a fair comparison, the user sum rate achieved in one round communication is normalized with respect to the number of time slots in each round. Thus the achievable sum rate for user 1 and user 2 is expressed as

$$R_i^{NC} = \frac{1}{2}\log_2(1+\gamma_i), i = \{u1, u2\}, \tag{4.12}$$

where the factor 1/2 is used to account for two time slots needed to complete one round transmission.

NOMA TDMA Scheme

NOMA TDMA scheme needs three time slots to complete one round communication. The first slot does the same as the first time slot in the NOMA cooperative scheme. Afterwards, relay 1 sends x_1 to user 1 in the second time slot while relay 2 sends x_2 to user 2 in the third time slot, as shown in Fig. 4.2.

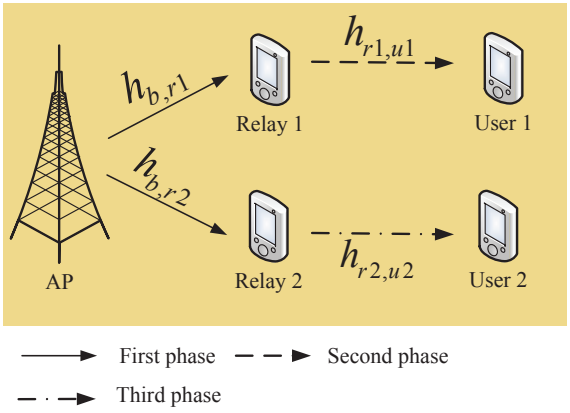

Fig. 4.2: NOMA TDMA scheme

By receiving messages separately in time slots 2 and 3, users 1 and 2 will not experience interference from each other. As a result, the achievable sum rate is

$$R_i^{NT} = \frac{1}{3}\log_2(1+\frac{|h_{j,i}|^2 P_r}{N_0}), \tag{4.13}$$

where $i = \{u1, u2\}, j = \{r1, r2\}$. Likewise, we use the factor $\frac{1}{3}$ to indicate three time slots in this scenario. Note that when calculating the first time slot data rate, we also need to use the factor 1/3, i.e.,

$$R_{i,j} = \frac{1}{3}\log_2(1 + \gamma_{i,j}), \tag{4.14}$$

where $i = \{r1, r2\}, j = \{x1, x2\}$.

4.1.3 Outage Probability Analysis

In this section, we analyze the system performance in terms of outage probability, which represents the probability of an event that the achieved data rate is less than a predefined one. Outage probability is a good metric for QoS in the system design. A closed-form analytical outage probability is derived for different users, based on which a high SNR approximation will also be presented.

Outage Probability in NOMA Cooperative Scheme

Let R_1 and R_2 denote the predefined minimum rates for user 1 and user 2 respectively. An outage occurs when the achievable data rate is less than the minimum data rate. Define $\mathcal{O}_{u1,NC}$ as the event of an outage at user 1. We first consider the complimentary event of $\mathcal{O}_{u1}^{NC}$, which is denoted as $\mathcal{O}_{u1,NC}^{C}$. The second time slot transmission relies on the successful decoding at the first time slot. For a DF relaying scheme, $\mathcal{O}_{u1,NC}^{C}$ happens when relay 1 successfully decodes x_1, and relay 2 successfully decodes x_2, and user 1 successfully decodes x_1. Thus, the outage probability can be calculated as

$$\begin{aligned} P(\mathcal{O}_{u1,NC}) &= 1 - P(\mathcal{O}_{u1,NC}^{C}) \\ &= 1 - P\Big(\min\{R_{r1,x1}, R_{u1}\} > R_1 \ \textit{and} \ \min\{R_{r1,x2}, R_{r2,x2}\} > R_2 \Big). \end{aligned} \tag{4.15}$$

Similarly, the outage probability for user 2 is

$$\begin{aligned} P(\mathcal{O}_{u2,NC}) &= 1 - P(\mathcal{O}_{u2,NC}^{C}) \\ &= 1 - P\Big(R_{r1,x1} > R_1 \ \textit{and} \ \min\{R_{r1,x2}, R_{r2,x2}, R_{u2}\} > R_2 \Big). \end{aligned} \tag{4.16}$$

Lemma 1. *([34] , Theorem 2.3.18) Let* $\mathbf{H}$ *be a* 2×2 *matrix and its entries follow i.i.d. Gaussian distribution with zero mean and unit variance. If* $\mathbf{H} = \mathbf{LQ}$*, where* $\mathbf{L}$ *is a lower triangle matrix and* $\mathbf{Q}$ *is a semi-orthogonal matrix, then* $|l_{1,1}|^2 \sim \chi^2(4)$ *and* $|l_{2,2}|^2 \sim \exp(1)$*.*

Theorem 1. *The outage probabilities for user 1 and user 2 in NOMA cooperative scheme can be expressed as*

$$P(\mathcal{O}_{u1,NC}) = 1 - e^{-\frac{\phi_1}{\sigma^2_{b,r1}}} e^{-\frac{\phi_2}{\sigma^2_{b,r2}}} (\phi_3 + 1) e^{-\phi_3} \tag{4.17}$$

and

$$P(\mathcal{O}_{u2,NC}) = 1 - e^{-\frac{\phi_1}{\sigma^2_{b,r1}}} e^{-\frac{\phi_2}{\sigma^2_{b,r2}}} e^{-\phi_4}, \tag{4.18}$$

where $\rho_s \triangleq \frac{P_s}{N_0}$, $\rho_r \triangleq \frac{P_r}{N_0}$, $z1 \triangleq 2^{2R_1} - 1$, $z2 \triangleq 2^{2R_2} - 1$, $\phi_1 = \max\{\frac{z_1}{\alpha_s \rho_s},\ \phi_2\}, \phi_2 = \frac{z_2}{(\beta_s - z_2\alpha_s)\rho_s}$, $\phi_3 = \frac{z_1}{\rho_r}$, $\phi_4 = \frac{z_2}{\rho_r}$.

Proof. From equation (4.15),

$$\begin{aligned}
P(\mathcal{O}_{u1,NC}) &= 1 - P\Big(\min\{R_{r1,x1}, R_{u1}\} > R_1 \ \textit{and}\ \min\{R_{r1,x2}, R_{r2,x2}\} > R_2 \Big) \\
&= 1 - P\Big(\min\{\frac{1}{2}\log_2(1+\gamma_{r1,x1}), \frac{1}{2}\log_2(1+\gamma_{u1})\} > R_1 \Big) * \\
&\quad P\Big(\min\{\frac{1}{2}\log_2(1+\gamma_{r1,x2}), \frac{1}{2}\log_2(1+\gamma_{r2,x2})\} > R_2 \Big) \\
&\overset{a}{=} 1 - P\big(|h_{b,r1}|^2 > \phi_1\big) P\big(|h_{b,r2}|^2 > \phi_2\big) P\big(|l_{1,1}|^2 > \phi_3\big) \\
&\overset{b}{=} 1 - e^{-\frac{\phi_1}{\sigma^2_{b,r1}}} e^{-\frac{\phi_2}{\sigma^2_{b,r2}}} (\phi_3 + 1) e^{-\phi_3}.
\end{aligned}$$

Here, $\overset{a}{=}$ holds when $\beta_s > \max\{z_2\alpha_s, \alpha_s\}$. $\overset{b}{=}$ holds since both $|h_{b,r1}|^2$ and $|h_{b,r2}|^2$ follow an exponential distribution with parameter 1 while $|l_{1,1}|^2$ follows a chi-squared distribution with a degree of freedom 4.

Similarly,

$$\begin{aligned}
P(\mathcal{O}_{u2,NC}) &= 1 - P\bigg(R_{r1,x1} > R_1 \ and \ \min\{R_{r1,x2}, R_{r2,x2}, R_{u2}\} > R_2\bigg) \\
&= 1 - P\bigg(\frac{1}{2}\log_2(1+\gamma_{r1,x1}) > R_1\bigg) * \\
&\quad P\bigg(\min\{\frac{1}{2}\log_2(1+\gamma_{r1,x2}), \frac{1}{2}\log_2(1+\gamma_{r2,x2}), \frac{1}{2}\log_2(1+\gamma_{u2})\} > R_2\bigg) \\
&= 1 - P\big(|h_{b,r1}|^2 > \phi_1\big)P\big(|h_{b,r2}|^2 > \phi_2\big)P\big(|l_{2,2}|^2 > \phi_4\big) \\
&= 1 - e^{-\frac{\phi_1}{\sigma^2_{b,r1}}} e^{-\frac{\phi_2}{\sigma^2_{b,r2}}} e^{-\phi_4}.
\end{aligned}$$

□

Since $\lim_{x\to 0}(1-e^{-x}) \simeq x$, in the high SNR regime, i.e., when $\rho_s, \rho_r \to \infty$, user 2 outrage probability at high SNR can br approximated as:

$$P(\mathcal{O}_{u2,NC}) = \frac{\phi_1}{\sigma^2_{b,r1}} + \frac{\phi_2}{\sigma^2_{b,r2}} + \phi_4. \tag{4.19}$$

4.1.4 Outage probability in NOMA TDMA scheme

As previously stated the first time slot in this scheme also uses NOMA transmission from the AP to two relays. Afterwards two relays will transmit x_1 and x_2 to the respective recipient in the following two time slots separately. Similar to NOMA cooperative scheme, the expressions for outage probabilities for user 1 and user 2 are respectively expressed as

$$P(\mathcal{O}_{u1,NT}) = 1 - P\bigg(\min\{R_{r1,x1}, R_{u1}\} > R_1 \ and \ R_{r1,x2} > R_2\bigg) \tag{4.20}$$

and

$$P(\mathcal{O}_{u2,NT}) = 1 - P\bigg(\min\{R_{r2,x2}, R_{u2}\} > R_2\bigg) \tag{4.21}$$

We have the following theorem for the outage probabilities.

Theorem 2. *The outage probabilities for user 1 and user 2 in NOMA TDMA scheme can be calculated as*

$$P(\mathcal{O}_{u1,NT}) = 1 - e^{-\frac{\phi_5}{\sigma^2_{b,r1}}} e^{-\phi_6} \tag{4.22}$$

and

$$P(\mathcal{O}_{u2,NT}) = 1 - e^{-\frac{\phi_7}{\sigma^2_{b,r2}}} e^{-\phi_8}, \tag{4.23}$$

where $\phi_5 = \max\{\frac{z_3}{\alpha_s \rho_s}, \phi_7\}$, $\phi_6 = \frac{z_3}{\rho_r}$, $\phi_7 = \frac{z_4}{(\beta_s - z_4\alpha_s)\rho_s}$, $\phi_8 = \frac{z_4}{\rho_r}$, $z_3 = 2^{3R_1} - 1$, *and* $z_4 = 2^{3R_2} - 1$.

The proof is similar to **Theorem 1** and thus is not detailed here. Note that in order for this theorem to hold, we need to have $\beta_s > \max\{z_4\alpha_s, \alpha_s\}$.

4.1.5 Outage Probability with Error Propagation in SIC

In previous sections, we have derived the outage performance for user 1 and user 2 by assuming both relays can decode NOMA signals correctly by using SIC. In what follows, we introduce the concept of error propagation in SIC, which can affect the system performance such as sum rate and outage probability.

The process of SIC consists of decoding, reconstruction and subtraction (DRS) [33]. Take relay 1 as an example, upon receiving the superimposed signal, x_2 will be decoded first by treating x_1 as interference. Then a reconstruction process will take place where relay 1 estimates its channel gain and uses the decoded signal $\hat{x}_2$. Therefore, the superposition signal for the next decoding symbol x_1 will be updated to

$$y_{r1,x1} = y_{b,r1} - \hat{h}_{b,r1}\hat{x}_2, \tag{4.24}$$

where $\hat{h}_{b,r1}$ is the estimated channel gain for relay 1. Existing papers assume the perfect decoding and cancellation of x_2 and thus we have $y_{r1,x1} = h_{b,r1}\sqrt{\alpha_s P_s}x_1 + n_{b,r1}$. We argue that this is a strong assumption since neither the channel estimation nor signal decoding can be perfect. While we desire to let $\hat{h}_k$ and $\hat{s}_M$ as close to h_k and s_M as possible, factors such as synchronization, phase ambiguity and deep fading can seriously degrade the SIC

process and errors can be accumulated and affect the UE to be decoded afterwards. We refer this process as EP.

θ is defined as the EP factor in this paper. Since there is a residual power when decoding the second signal, (4.4) and (4.6) can be updated to,

$$\gamma_{r1,x1}^{EP} = \frac{\alpha_s P_s |h_{b,r1}|^2}{N_0 + \theta\beta_s P_s |h_{b,r1}|^2}, \qquad \gamma_{r2,x1}^{EP} = \frac{\alpha_s P_s |h_{b,r2}|^2}{N_0 + \theta\beta_s P_s |h_{b,r2}|^2}. \tag{4.25}$$

θ represents the amount of residual power from the previous decoding and $0 \leq \theta \leq 1$. When $\theta = 0$, the results agree with perfect cancellation. $\theta = 1$ is the worst case when SIC fails to decode the first signal and the second stage decoding has to treat the entire first signal as interference. Besides, θ should be inversely proportional to the SNR of x_2. In this paper, we assume θ is a constant for simplicity.

Similarly, the outage probability analysis is given as follows.

Outage Probability in NOMA Cooperative Scheme with EP

Define $\mathcal{O}_{i,NC}^{EP}, i = \{u1, u2\}$ as the outage event of user i in the NOMA cooperative scheme. Then, we have the following theorem for the outage probability.

Theorem 3. *The outage probabilities for user 1 and user 2 in the NOMA cooperative scheme when considering EP in SIC are respectively derived as*

$$P(\mathcal{O}_{u1,NC}^{EP}) = 1 - e^{-\max\{\frac{z1}{(\alpha_s - z_1\theta\beta_s)\rho_s}, \phi_2\}/\sigma_{b,r1}^2} e^{-\phi_2/\sigma_{b,r2}^2} (\phi_3 + 1) e^{-\phi_3} \tag{4.26}$$

and

$$P(\mathcal{O}_{u2,NC}^{EP}) = 1 - e^{-\max\{\frac{z1}{(\alpha_s - z_1\theta\beta_s)\rho_s}, \phi_2\}/\sigma_{b,r1}^2} e^{-\phi_2/\sigma_{b,r2}^2} e^{-\phi_4}. \tag{4.27}$$

Outage Probability in NOMA TDMA Scheme with EP

Similarly, we also consider EP in SIC in the second scheme. Let $\mathcal{O}_{i,NT}^{EP}, i = \{u1, u2\}$ be the event of an outage. We have the following theorem for the analytical results of outage

probabilities.

Theorem 4. *The outage probabilities for user 1 and user 2 in NOMA TDMA transmission when considering EP in SIC are*

$$P(\mathcal{O}_{u1,NT}^{EP}) = 1 - e^{-\max\{\frac{z3}{(\alpha_s - z_3\theta\beta_s)\rho_s}, \phi_7\}/\sigma_{b,r1}^2} e^{-\phi_6} \tag{4.28}$$

and

$$P(\mathcal{O}_{u2,NT}^{EP}) = 1 - e^{-\frac{\phi_7}{\sigma_{b,r2}^2}} e^{-\phi_8}. \tag{4.29}$$

Remark 1: The constraint for **Theorem 3** to hold is $\frac{\alpha_s}{z_1\theta} > \beta_s > \max\{z_2\alpha_s, \alpha_s\}$, which has one additional constraint ($\alpha_s > z_1\theta\beta_s$) compared with **Theorem 1**. Similarly, **Theorem 4** holds when $\frac{\alpha_s}{z_3\theta} > \beta_s > \max\{z_4\alpha_s, \alpha_s\}$, which also posts another constraint. These additional constraints can potentially increase the outage probability.

Remark 2: We show that power allocation factors impact the outage probability. Specifically, if the constraints in *Remark 1* cannot be satisfied, the outage probability will always be 1 for both users, which indicates the failure of both schemes. The reason is that if x_2 cannot be decoded in the first time slot, then the second or third time slot can not proceed. If the maximum value of β_s is less than $\frac{\alpha_s}{z_1}\Omega_1$, where $\Omega_1 = \min\{z_2(1+z_1), \frac{1}{\theta}, \frac{z_2(1+z_1)}{1+z_2\theta}\}$, the outage probability under NOMA cooperative scheme is the same for both with EP or without EP cases. Likewise, for NOMA TDMA scheme, the outage probability is the same for user 1 in both EP or no EP cases when $\beta_s < \frac{\alpha_s}{z_3}\Omega_2$, where $\Omega_2 = \min\{z_4(1+z_3), \frac{1}{\theta}, \frac{z_4(1+z_3)}{1+z_4\theta}\}$. The reason is that if we limit the value of β_s, the bottleneck of the data rate does not come from the first time slot transmission, which may be directly affected by the EP. However, for user 2, the outage probability is always identical with or with not EP since under this circumstance user 2 is not impacted by EP.

4.1.6 Performance Study

In this section, performance evaluation on the proposed schemes are provided based on both simulation and analysis. Some basic parameters are set as follows. The channel gains

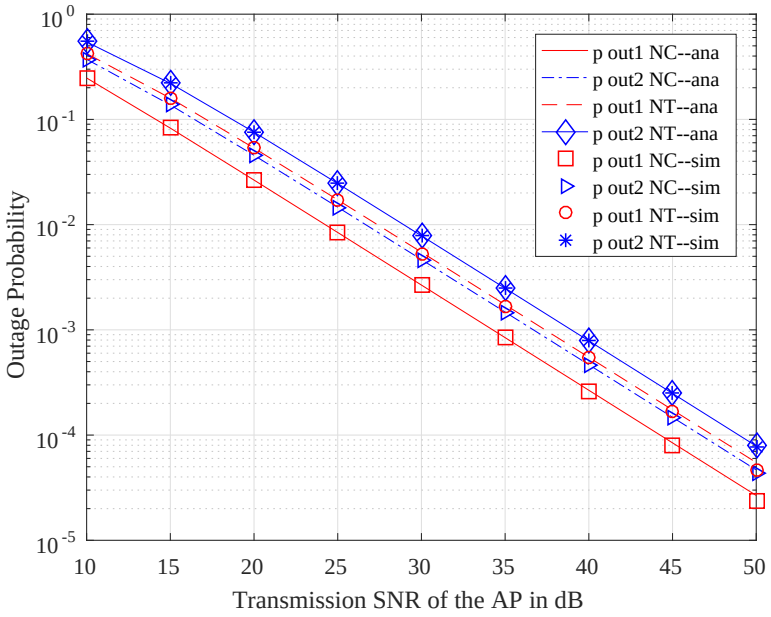

Fig. 4.3: Theorem 1 and 2. $\alpha_s = 0.2, \beta_s = 0.8. R_1 = R_2 = 0.5$ bps/Hz.

of $h_{b,r1}$ and $h_{b,r2}$ are 5 and 1 respectively, i.e., $\sigma^2_{b,r1} = 5, \sigma^2_{b,r2} = 1$. The transmission SNR of the AP ranges from 10 dB to 50 dB, and the transmit power of both relays is set to half of the AP's power, which means there is a 3 dB difference between P_s and P_r.

Fig. 4.3 illustrates the outage performance in both schemes with perfect SIC, i.e., no EP in SIC, as a function of the AP transmit SNR in dB. The predefined minimum data rates R_1 and R_2 are both set to 0.5 bps/Hz. Besides, $\alpha_s = 0.2$ and $\beta_s = 0.8$ are constraints. Apparently, optimizing α_s and β_s based on channel condition and transmit SNR will further improve the outage probability performance and this can be explored in the future work. It is observed that all the outage probabilities decrease with the increment of SNR. The analytical results match the simulation results very well, which validates the earlier analysis in **Theorem. 1** and **Theorem. 2**. Because of this, we only present the analytical results for better illustrations in the figures for the following parts.

Further, by comparing the performance of NOMA cooperative and NOMA TDMA schemes, one can conclude that NOMA cooperative scheme achieves lower outage probabilities than the NOMA TDMA scheme, which uses three time slots in one round communi-

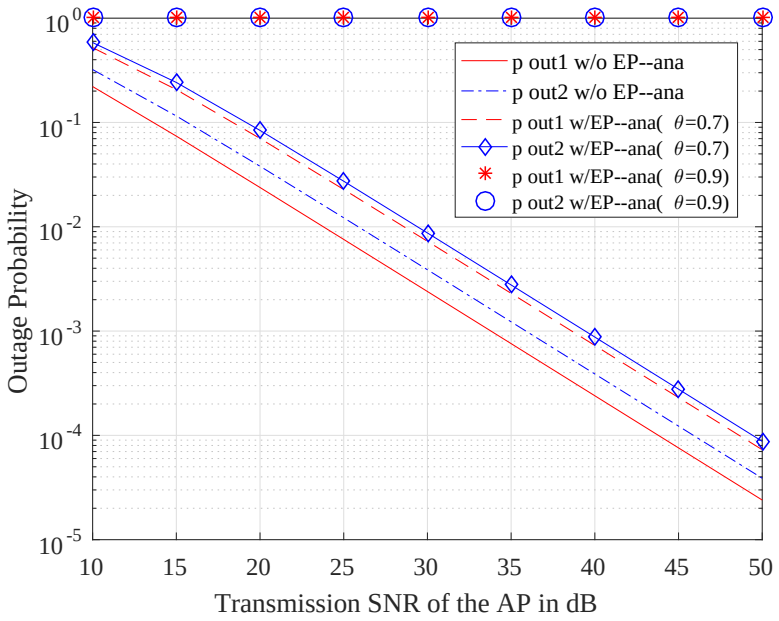

Fig. 4.4: Theorem 3. $\alpha_s = 0.36, \beta_s = 0.64. R_1 = R_2 = 0.4$ bps/Hz. $\theta = 0.7$ and $\theta = 0.9$.

cation and hence the added factor $\frac{1}{3}$ decreases the total sum rate. In both schemes, user 1 outperforms user 2 since user 2's message x_2 is decoded first, which has a higher interference term.

Fig. 4.4 presents the result for **Theorem. 3**. A new set of parameters is selected to satisfy *Remark. 1* and *Remark. 2*. The corresponding parameters are $\alpha_s = 0.36, \beta_s = 0.64, R_1 = R_2 = 0.4$ bps/Hz. The curve without EP is also plotted for reference. One can see that SIC EP degrades the outage performance largely when $\theta = 0.7$. However, when $\theta = 0.9$, the condition $\frac{\alpha_s}{z_1\theta} > \beta_s$ is not satisfied any more, making both the analytical and simulated outage probabilities to become 1.

The result for **Theorem. 4** is shown in Fig. 4.5. Likewise, we plot the case without EP for reference. The parameters for this scheme are $\alpha_s = 0.36, \beta_s = 0.64, R_1 = R_2 = 0.4$ bps/Hz. These parameters are selected to meet the requirements of *Remark. 1* and *Remark. 2*. When EP is considered, the performance becomes worse for user 1 while user 2 outage probability remains the same . When $\theta = 0.6$, the condition $\frac{\alpha_s}{z_1\theta} > \beta_s$ is not satisfied. As expected, the outage probability becomes 1 for user 1.

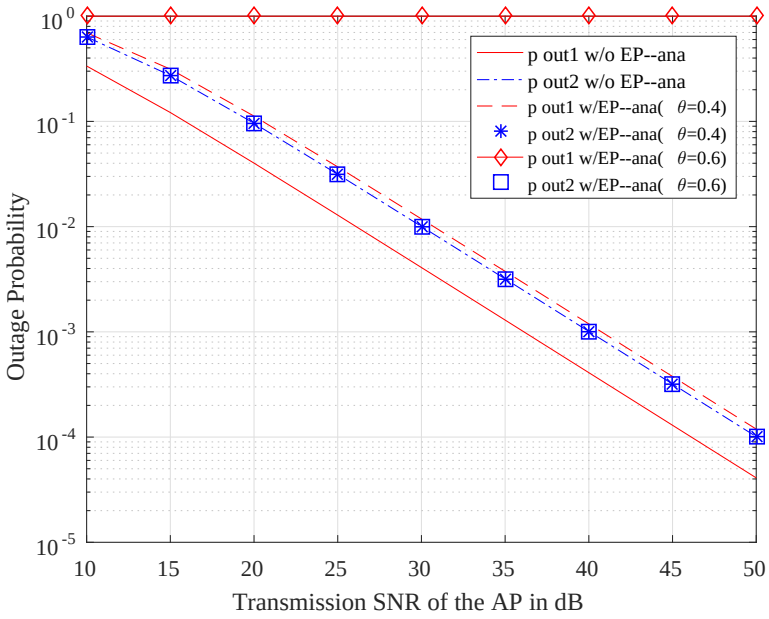

Fig. 4.5: Theorem 4. $\alpha_s = 0.36, \beta_s = 0.64. R_1 = R_2 = 0.4$ bps/Hz. $\theta = 0.4$ and $\theta = 0.6$.

4.2 Non-Orthogonal Multiple Access in a mmWave Based IoT Wireless System with SWIPT

4.2.1 Introduction

The unprecedented growth of mobile devices including smart phones, tablets, laptops, and IoT devices drives the wireless telecommunication industry to a new level. The requirements come from various aspects such as higher data rate, fairness, tremendous connectivity, and low latency from different applications and various end users. Therefore, as a new generation technology, 5G emerges with its goal to provide 1000 times higher data rate, 1 ms low latency, and support billions of upcoming IoT devices. Among these features, 1000 times capacity can be achieved by the new mmWave spectrum, novel network architectures and new radio access technologies (RATs) [27].

Due to the ad hoc deployment nature of most low-power nodes and devices, they may have limited access to wireline power charging facilities and also have limited battery life. In this paper, low-power relay nodes and devices are assumed to be capable of energy harvest

functionality. More specifically, SWIPT is considered. SWIPT can have two implementation modes, namely time switching (TS) mode and power splitting (PS) mode [35]. In the TS mode, a dedicated resource is used for energy transfer from which the harvested energy is then used for future information transmission. In the PS mode, upon receiving the radio signal, the energy harvest node splits the signal into two parts. The first part is used for signal decoding while the second part is used for energy charging. A linear energy harvest model, which assumes the output power of the energy harvest circuit grows linearly with the input power, is applied in most existing works. Cooperative NOMA system with SWIPT is studied in [59], where they proposed different user selection schemes and evaluate the performance with outage probability. This paradigm is proved impractical based on field test results as shown in [36]. As a result, a more practical yet more complicated non-linear model which better matches current circuit design is considered in this paper. Thus the wireless heterogeneous system in this study consists of higher power MBSs and low-power relays with SWIPT that is based on the non-linear energy harvesting model. Downlink NOMA is first used to transmit composite signals to UE and relay. Relay then harvests the energy by using non-linear model in PS mode. With the harvested energy, relay sends the received signal to the cell edge UE.

4.2.2 System Model

The system model is based on a mmWave downlink wireless heterogeneous system that consists of high power MBSs, low-power relays, and low-power IoT devices, such as sensors or wearable devices. At mmwave band, MBSs are equipped with a large number of horn antennas, which have narrow half-power-beamwidth (HPBW) to combat with the severe pathloss and each transmission is conducted with a single antenna. While each low-power relay or IoT device is equipped with a single antenna due to the size and power constraints. It is assumed that MBSs can coordinate the transmission direction with a stepper motor, hence inter-cell and intra cell interference can be eliminated by carefully aligning the beam directions. Furthermore, relaying and NOMA are used to help reach UEs out of coverage due to severe blockage at mmWave band. Without loss of generality, IoT UE 1 and IoT

UE 2 are selected, where UE 1 is in the beamforming coverage area while there is a severe blockage between BS and UE2 so that a direct transmission link between the MBS and UE 2 is difficult to establish. Thus BS can communicating to UE 2 through relays. In this paper we assume D2D relaying mode is used so that the relay can communication with a UE in close proximity and we assume the relay is capable of rechargeable functionality. So the power consumed for relaying comes directly from electromagnetic waves, which can relieve the concern on limited battery life for typical IoT devices. With NOMA and relay, complete transmission cycle consists of two phases. In the first phase, the BS sends a composite signal to UE 1 and a selected relay device simultaneously by applying NOMA. After receiving the signal, the relay device splits the signal into two parts. One part is for information decoding and the other part is for energy harvesting. In the second phase, the BS sends another message to UE 1 while the relay device sends the decoded message to UE 2 by using the harvested energy in phase 1.

Denote the channel between BS and UE 1, BS and relay device, relay device and UE 2 as h'_{B1}, h'_{BR}, and h'_{R2}, respectively. Frequency flat quasi-static block fading model is used here so the channel does not change during the two transmission phases while the channel changes from cycle to cycle. Additionally, $h'_i = \frac{h_i\sqrt{a_0}}{\sqrt{1+d_i^\alpha}}$, where h_i is modeled as Rayleigh fading with $h_i \sim \mathcal{CN}(0,1), i = \{B1, BR, R2\}$ [37]. a_0 is antenna-specific gain for the BS and $a_0 = 1$ when $i = R2$. An illustration of the system model is in Fig. 4.6. In the following, the transmission process for each cycle is illustrated.

Phase 1 Transmission

In this phase, the BS sends the superimposed message to both UE 1 and the relay. The message is given as $x = \sqrt{\lambda_1 P_{BS}}x_1 + \sqrt{\lambda_2 P_{BS}}x_2$, where λ_1 and λ_2 are power allocation factors for UE 1 and the relay respectively with $\lambda_1 + \lambda_2 = 1$. x_1 and x_2 are normalized intended signal for UE 1 and UE 2. P_{BS} is the transmission power of the BS. At the receiver

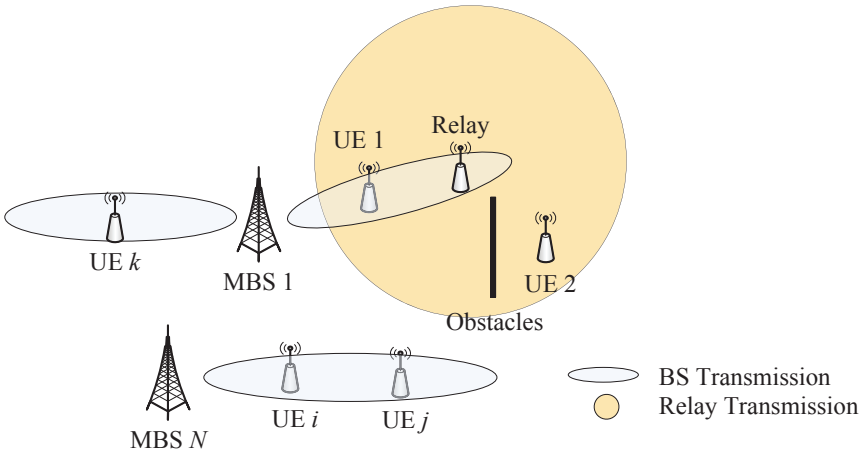

Fig. 4.6: System model

side, UE 1 observes y^1_{UE1}, which is expressed as

$$\begin{aligned} y^1_{UE1} &= \frac{h_{B1}\sqrt{a_0}}{\sqrt{1+d^{\alpha}_{B1}}}x + n_{B1} \\ &= \frac{h_{B1}\sqrt{a_0}}{\sqrt{1+d^{\alpha}_{B1}}}(\sqrt{\lambda_1 P_{BS}}x_1 + \sqrt{\lambda_2 P_{BS}}x_2) + n_{B1}, \end{aligned} \tag{4.30}$$

where n_{B1} is the additive Gaussian white noise (AWGN) with variance σ^2, d_{B1} is the distance from the BS to UE 1, α is the path loss exponent for line-of-sight (LOS).

Without loss of generality, we assume $|h_{B1}|^2 > |h_{BR}|^2$. Hence according to NOMA protocol, $\lambda_1 < \lambda_2$ is set to ensure QoS at the weak receiver. With this setting, UE 1 first decodes signal x_2 with its SINR formulated as

$$\gamma^1_{UE1,x_2} = \frac{\lambda_2 \rho_{B1}|h_{B1}|^2}{\lambda_1 \rho_{B1}|h_{B1}|^2 + 1}, \tag{4.31}$$

where $\rho_{B1} = \frac{P_{BS}a_0}{\sigma^2(1+d^{\alpha}_{B1})}$ is the transmission SNR from the BS to UE 1. The superscript "1" indicates the first phase. SIC is performed to remove x_2 from the superimposed signal, then UE 1 can decode its own message with the following SINR

$$\gamma^1_{UE1,x_1} = \lambda_1 \rho_{B1}|h_{B1}|^2. \tag{4.32}$$

At the relay side, it first splits the observation into two parts. One part is for the rechargeable unit, which consists of a super capacitor or a short-term high efficiency battery. The other part is for information decoding, which can be expressed as

$$\begin{aligned} y_R^D &= \frac{h_{BR}\sqrt{a_0}}{\sqrt{1+d_{BR}^{\alpha}}} x\sqrt{1-\beta} + n_{BR} \\ &= \frac{h_{BR}\sqrt{a_0}}{\sqrt{1+d_{BR}^{\alpha}}} \sqrt{1-\beta}(\sqrt{\lambda_1 P_{BS}} x_1 + \sqrt{\lambda_2 P_{BS}} x_2) + n_{BR}, \end{aligned} \tag{4.33}$$

where β is the power split coefficient indicating the portion of power assigned to energy harvest unit. n_{BR} has the same distribution with n_{B1}. Signal y_R^D goes through the decoding unit for x_2, the corresponding SINR is

$$\gamma_{R,x_2}^1 = \frac{(1-\beta)\lambda_2 \rho_{BR} |h_{BR}|^2}{(1-\beta)\lambda_1 \rho_{BR} |h_{BR}|^2 + 1}, \tag{4.34}$$

where $\rho_{BR} = \frac{P_{BS} a_0}{\sigma^2 (1+d_{BR}^{\alpha})}$ is the transmission SNR from the BS to the relay.

The remaining power $P_R^C = |h_{BR}|^2 \beta \rho_{BR} \sigma^2$ is harvested by the relay. In this paper, we adopt the non-linear energy harvest model, which is more precise in modeling the power-in-power-out relation in current wireless charging technology. Specifically, the harvested energy can be expressed as a logistic (sigmoidal) function

$$P_R^{EH} = \frac{M}{1 + \exp\big(-a(P_R^C - b)\big)}, \tag{4.35}$$

where M, a, b are constants and represent different physical meanings in wireless charging. M denotes the maximum harvested power at the relay when the energy harvesting circuit is saturated. a together with b capture the joint effect of resistance, capacitance and circuit sensitivity [38].

[36] provides a more sophisticated model, which captures the zero-input-zero-output feature in wireless charging and can be modeled in the following.

$$P_R^{EH} = \frac{\Psi - M\Omega}{1-\Omega}, \quad \Omega = \frac{1}{1+\exp(ab)}, \tag{4.36}$$

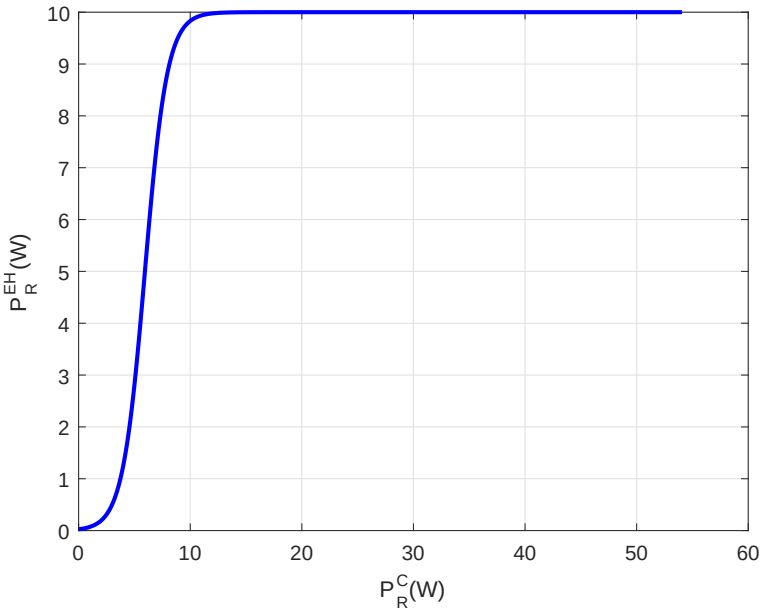

Fig. 4.7: Power-in-power-out response in the non-linear energy harvest model

where $\Psi = \frac{M}{1+\exp\left(-a(P_R^C - b)\right)}$.

In the subsequent analysis, we use model (4.35) based on the following reasons. 1) Our model does not have zero power input case; 2) The general logistic function can reduce the complexity in outage analysis; 3) (4.35) can provide sufficient precision.

Fig. 4.7 presents the power-in-power-out relation with 1000 independent events, based on which the parameters are estimated as follows, $\beta = 0.6$, $\sigma = 0.0995$, $M = 10$, $a = 1$, $b = \beta\rho_{BR}\sigma^2$, and $\rho_{BR} = 30$ dB.

Phase 2 Transmission

During the second phase, the relay sends x_2 to UE 2 with the energy harvested in *Phase 1*. Meanwhile, the BS sends another signal x_3 to UE 1. The received signal at UE 1 and UE 2 are expressed as

$$y_{UE1}^2 = \sqrt{P_{BS}}\frac{h_{B1}\sqrt{a_0}}{\sqrt{1+d_{B1}^{\alpha}}}x_3 + \sqrt{P_R^{EH}}\frac{h_{R1}}{\sqrt{1+d_{R1}^{\alpha}}}x_2 + n_{B1} \tag{4.37}$$

and

$$y_{UE2}^2 = \sqrt{P_R^{EH}} \frac{h_{R2}}{\sqrt{1+d_{R2}^{\alpha}}} x_2 + n_{B2}, \tag{4.38}$$

respectively.

Since UE 1 already decodes x_2 in *Phase 1*, by appropriately estimating the channel h_{R1}, it can employ SIC to subtract x_2 from its observation [31]. The remaining SINR becomes

$$\gamma_{UE1,x_3}^2 = \rho_{B1}|h_{B1}|^2. \tag{4.39}$$

For UE 2, since there is a severe blockage between BS and itself, it has a negligible interference from BS. The achievable SINR at UE 2 is

$$\gamma_{UE2.x_2}^2 = \rho_{EH}|h_{R2}|^2, \tag{4.40}$$

with $\rho_{EH} = \frac{P_R^{EH}}{\sigma^2(1+d_{R2}^{\alpha})}$.

4.2.3 Outage Analysis

In this section, we will provide mathematical analysis on the outage probability of the proposed scheme. The outage probability is defined as the probability of events where certain measurements such as SINR or data rate cannot meet the pre-defined threshold.

UE 1 Outage Probability

Define the minimum data rates for for messages x_1, x_2 and x_3 as R_1, R_2 and R_3 respectively. Below the minimum data rate, a UE will have an outage. Since UE 1 involves in both phases, outage occurs when UE 1 fails to decode x_2 and x_1 in phase 1 or fails to decode x_3 in phase 2. For simplicity, we can consider the complementary event first. Specifically, we can derive the outage probability of UE 1 as follows.

$$
\begin{aligned}
P(\mathcal{O}_{UE1}) &= 1 - P(\mathcal{O}_{UE1}^{C}) \\
&= 1 - P\bigg(\frac{1}{2}\log_2(1+\gamma_{UE1,x_2}^{1}) > R_2 \ \textit{and}\ \frac{1}{2}\log_2(1+\gamma_{UE1,x_1}^{1}) > R_1 \\
&\quad \textit{and}\ \frac{1}{2}\log_2(1+\gamma_{UE1,x_3}^{2}) > R_3\bigg).
\end{aligned}
$$

Notice that channel $h_{B1} \sim \mathcal{CN}(0,1)$ and $|h_{B1}|^2 \sim \exp(1)$. Define $z_1 = 2^{2R_1}-1$, $z_2 = 2^{2R_2}-1$ and $z_3 = 2^{2R_3}-1$.

$$
\begin{aligned}
P(\mathcal{O}_{UE1}) &= P(|h_{B1}|^2 > \phi_1) \\
&= 1 - e^{-\phi_1},
\end{aligned}
\tag{4.41}
$$

where $\phi_1 = \max\{\frac{z_2}{\lambda_2\rho_{B1} - z_2\lambda_1\rho_{B1}}, \frac{z_1}{\lambda_1\rho_{B1}}, \frac{z_3}{\rho_{B1}}\}$.

Note that the above outage probability is conditioned on $\lambda_2 > z_2\lambda_1$. Otherwise the outage occurs with probability 1.

UE 2 Outage Probability

For UE 2, since the BS only transmits x_2 via the relay. Thus the bottleneck of this transmission depends on the minimum data rate in two phases. The outage probability for UE 2 is

$$
\begin{aligned}
P(\mathcal{O}_{UE2}) &= 1 - P(\mathcal{O}_{UE2}^{C}) \\
&= 1 - P\bigg(\min\{\frac{1}{2}\log(1+\gamma_{R,x_2}^{1}), \frac{1}{2}\log(1+\gamma_{UE2,x_2}^{2})\} > R_2\bigg) \\
&= 1 - P\bigg(\min\{\gamma_{R,x_2}^{1}, \gamma_{UE2,x_2}^{2})\} > z_2\}\bigg).
\end{aligned}
\tag{4.42}
$$

The following theorem provides an analytical result for the outage probability of UE 2.

Theorem 5. *The outage probability for UE 2 in the proposed non-linear energy harvest model is* $P(\mathcal{O}_{UE2}) = 1 - \frac{c_2}{c_4}e^{-c_1}(c_3e^{-c_1c_4})^{-\frac{1}{c_4}}\Gamma(\frac{1}{c_4}, c_3e^{-c_1c_4})$, *where* c_1, c_2, c_3 *and* c_4 *are constants and defined in the following proof.*

Proof. Let $c = (1 + d_{R2}^{\alpha})$, the outage probability becomes

$$\begin{aligned} P(\mathcal{O}_{UE2}) &= 1 - P\big(\min\{\gamma_{R,x_2}^1, \gamma_{UE2,x_2}^2\} > z_2\big) \\ &= 1 - P\big(\gamma_{R,x_2}^1 > z_2, \gamma_{UE2,x_2}^2 > z_2\big). \end{aligned} \tag{4.43}$$

Let probability $P\big(\gamma_{R,x_2}^1 > z_2, \gamma_{UE2,x_2}^2 > z_2\big)$ be P_1 for conciseness. Furthermore, let $|h_{BR}|^2 = x$ and $|h_{R2}|^2 = y$. x and y both follow an exponential distribution with parameter 1 and they are independent to each other.

$$\begin{aligned} P_1 &= P\Big(\frac{(1-\beta)\lambda_2\rho_{BR}x}{(1-\beta)\lambda_1\rho_{BR}x + 1} > z_2, \frac{P_R^{EH}}{\sigma^2 c} y > z_2\Big) \\ &\overset{a}{=} P\Big(x > \frac{z_2}{(1-\beta)\rho_{BR}(\lambda_2 - \lambda_1 z_2)}, \\ &\quad \frac{M}{\sigma^2 c(1 + \exp(-a(\beta\rho_{BR}\sigma^2 x - b)))} y > z_2\Big), \end{aligned} \tag{4.44}$$

where $\overset{a}{=}$ is conditioned on $\lambda_2 > \lambda_1 z_2$. Otherwise the outage probability will be always equal to one, as already observed in the existing literature. Define $f(x) = \frac{M}{\sigma^2 c(1+\exp(-a(\beta\rho_{BR}\sigma^2 x - b)))}$ and let $c_1 = \frac{z_2}{(1-\beta)\rho_{BR}(\lambda_2 - \lambda_1 z_2)}$. The above joint probability can be evaluated as

$$\begin{aligned} P_1 &= \int_{c_1}^{\infty} \int_{\frac{z_2}{f(x)}}^{\infty} e^{-x} e^{-y} dx dy \\ &= \int_{c_1}^{\infty} \exp\big(-x - \frac{z_2}{f(x)}\big) dx. \\ &= e^{-\frac{z_2\sigma^2 c}{M}} \int_{c_1}^{\infty} \exp\Big(-x - \frac{z_2\sigma^2 c}{M} e^{ab} \exp(-a\beta\rho_{BR}\sigma^2 x)\Big) dx. \end{aligned} \tag{4.45}$$

For notation simplicity, define $c_2 = e^{-\frac{z_2\sigma^2 c}{M}}$, $c_3 = \frac{z_2\sigma^2 c}{M} e^{ab}$ and $c_4 = a\beta\rho_{BR}\sigma^2$. Then P_1 can be simplified as

$$P_1 = c_2 \int_{c_1}^{\infty} \exp(-c_3 e^{-c_4 x} - x) dx. \tag{4.46}$$

Let $u = xc_4 - c_1c_4, u \in [0, \infty]$. According to ([39], 3.331-1)

$$\begin{aligned} P_1 &= \frac{c_2}{c_4} e^{-c_1} \int_0^{\infty} \exp(-c_3 e^{-c_1 c_4} e^{-u} - \frac{u}{c_4}) du \\ &= \frac{c_2}{c_4} e^{-c_1} (c_3 e^{-c_1 c_4})^{-\frac{1}{c_4}} \Gamma(\frac{1}{c_4}, c_3 e^{-c_1 c_4}). \end{aligned} \tag{4.47}$$

$\Gamma(\mu_2, \mu_1)$ is the lower incomplete gamma function, which is

$$\Gamma(\mu_2, \mu_1) = \int_0^{\mu_1} e^{-t} t^{\mu_2 - 1} dt, \tag{4.48}$$

where $\mu_2 > 0$.

□

Outage at High SNR

In this subsection, we provide the approximation for the outage probability at high SNR region. Specifically, if $\rho_{B1}, \rho_{BR} \to \infty$, the outage probability for UE 1 becomes

$$P(\mathcal{O}_{UE1}^H) = \phi_1 = \max\{\frac{z_2}{\lambda_2 \rho_{B1} - z_2 \lambda_1 \rho_{B1}}, \frac{z_1}{\lambda_1 \rho_{B1}}, \frac{z_3}{\rho_{B1}}\}, \tag{4.49}$$

since $\lim_{x \to 0}(1 - e^{-x}) \simeq x$.

For UE 2, the maximum charging power is M even when P_R^C becomes infinity. Thus, the high approximation becomes

$$P(\mathcal{O}_{UE2}^H) = 1 - P(\frac{\lambda_2}{\lambda_1} > z_2, \frac{M}{\sigma^2(1 + d_{R_2}^{\alpha})} |h_{R2}|^2 > z_2). \tag{4.50}$$

When $\frac{\lambda_2}{\lambda_1} > z_2$, the result becomes

$$P(\mathcal{O}_{UE2}^H) = 1 - e^{-\frac{z_2 \sigma^2 (1 + d_{R_2}^{\alpha})}{M}}. \tag{4.51}$$

Otherwise, if $\frac{\lambda_2}{\lambda_1} < z_2$, the outage probability will be always one in the high SNR regime.

Diversity Analysis for UE 2

Based on the definition of diversity, we have

$$d_{UE2} = -\lim_{\rho_{BR}\to\infty} \frac{\log P(\mathcal{O}_{UE2})}{\log \rho_{BR}} = 0. \tag{4.52}$$

This means in the non-linear energy harvest model, no diversity will be achieved. The reason is that as the input power increases, the power harvested becomes saturated, which limits the further data rate growth hence the outage probability performance.

4.2.4 Numerical Performance Results

In this section, numerical performance results are presented based on both simulations and analysis. The parameters for evaluation are chosen in the following. $a_0 = 4$, which indicates the horn antenna gain is 6 dB. $\lambda_1 = 0.4$, $\lambda_2 = 0.6$. $M = 4$, which means the maximum charging power for the relay is 4 Watts. For illustration purposes, the distance d_{BR}, d_{R2} and d_{B1} are small, which are set to 8, 2, and 10, respectively. Similar settings can also be found in [35]. Furthermore, the predefined thresholds for data rates are $R1 = R3 = 0.5$ bps/Hz and $R2 = 0.3$ bps/Hz.

Fig. 4.8 shows the outage probability of UE 1 and UE 2 with regards of the transmission SNR in dB. "ana" stands for analytical result while "sim" is the simulation one. The performance can be optimized by carefully choosing λ_1 and λ_2. The detailed study on how to select λ_1 and λ_2 values to achieve optimal performance is not the focus of this paper and hence not extended. Further, since a and b can also impact the system performance, the outage probability of UE 2 is evaluated with different a, b values. By fixing $\beta = 0.8$, both the simulation and analytical results are presented. As we can see from Fig. 4.8, the analytical results match well with the simulation ones for UE 1. As expected, the outage probability decreases linearly in log scale with the increase of transmission SNR. For UE 2, when $a = 2.5, b = 3$, the outage probability of UE 2 is lower than the case with $a = 6.5, b = 4$, which indicates that energy harvest circuit will affect the system performance. Also, as the transmission SNR becomes larger, the gap becomes less apparent. The reason

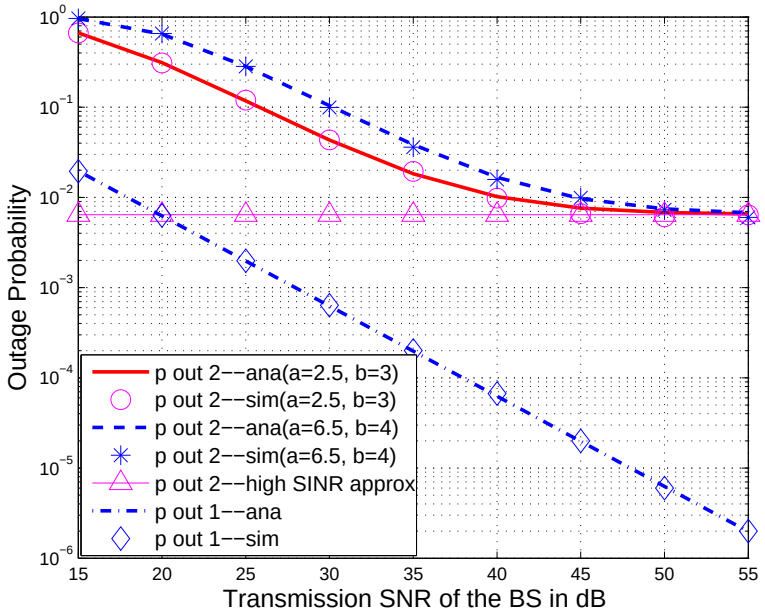

Fig. 4.8: Outage performance for both UEs with comparison to analytical results

is as SNR becomes larger, the harvested energy becomes a constant M, thus the outage performance becomes the same regarding different a and b values, as shown in the high SNR approximation part. Note that the non-linear response will only make sure the harvested energy does not exceed M. In some rare occasions, we can have $P_R^C < P_R^{EH}$, which clearly violates the physical meaning in our model. So these events are excluded from the results.

The outage performance for UE 2 as the function of β is shown in Fig. 4.9. The parameters used for this study are $a = 2, \rho_{BR} = 40$ dB. The simulation and analytical results for UE 2 are both presented here and they match well with each other. With the increase of β, the outage probability also increases. The increase slope slows down as β further increases, due to the fact that β is the portion of power assigned to energy harvest unit. The less power remained for transmitting, the higher outage probability it will have. The inconsistence between simulation and analytical results when $\beta = 0.1$ comes from the excluded events when $P_R^C < P_R^{EH}$.

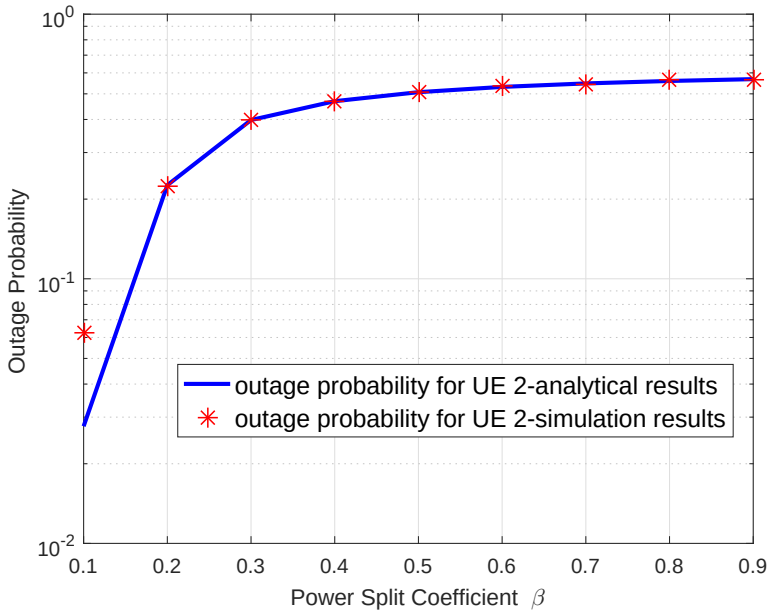

Fig. 4.9: Outage performance for UE 2 as the function of β

4.2.5 Chapter Conclusions

In the first part of the chapter, we analyze the outage performance of two NOMA relaying schemes. NOMA cooperative scheme needs two time slots to complete one round communication. It uses NOMA in the first time slot and uses DPC precoding in the second time slot for cooperation. NOMA TDMA scheme needs three time slots to complete one round communication. It uses NOMA in the first time slots and then TDMA in the second and third time slots. SIC Error propagation is considered in the analysis and the performance degradation is evaluated. The analytical results agree with the simulation results very well. Future work can optimize the power allocation factor α_s and β_s to achieve the best outage performance under different schemes.

In the second part, we consider applying NOMA and D2D relaying in a mmWave based wireless system that consists of high power base stations and low power IoT devices. The lower power IoT devices do not have external power supplies and have limited battery life. In order to prolong battery life and also to motivate low power IoT devices to help relay signals from others, low power IoT devices can harvest energy from electromagnetic signals.

To make the energy harvest model more realistic, non-linear energy harvesting model is used. The theoretical analysis on outage probability is given for the proposed scheme and system model. Simulation results validate the accuracy of the analysis.

CHAPTER 5

Robust Beamforming Design in a NOMA Cognitive Radio Network Relying on SWIPT

5.1 Introduction

As one promising technique of improving the SE, CR techniques have also been investigated for decades, where the secondary users (SUs) may access the spectrum bands of the primary users (PUs), as long as the interference caused by SUs is tolerable [44]. According to [45], in order to implement CR in practice, three operational models have been proposed, namely, opportunistic spectrum access, spectrum sharing, and sensing-based enhanced spectrum sharing. It is envisioned that the combination of NOMA with CR is capable of further improving the SE. As a benefit of its low implementational complexity, spectrum sharing has been widely applied. In [46]- [48], the authors analyzed the performance of a spectrum sharing CR combined with NOMA. It was shown that the SE can be significantly improved by using NOMA in CR compared to that achieved by using OMA in CR.

On the other hand, the increasing greenhouse gas emissions have become a major concern also in the design of wireless communication networks. According to [49], cellular networks world-wide consume approximately 60 billion kWh energy per year. Moreover, this energy consumption is explosively increasing due to the unprecedented expansion of wireless networks to support ubiquitous coverage and connectivity. Furthermore, because of the rapid proliferation of IoT applications, most battery driven power limited IoT devices become useless if their battery power is depleted. Thus it is critical to use energy in an efficient way or to harness renewable energy sources. As remedy, energy harvesting (EH) exploits the pervasive frequency radio signals for replenishing the batteries [50]. There have been two research thrusts on EH using RF technology. One focuses on wirelessly powered networks, where a so-called harvest-then-transmit protocol is applied [51]. The other one uses SWIPT [53]- [55], which is the focus of this chapter. The contributions of SWIPT

in CR has been extensively studied. Specifically, authors of [56] considered the optimal beamforming design in a multiple-input single-output (MISO) CR downlink network. A similar power splitting structure to that of our work is applied at the user side. Hu *et. al* [57], on the other hand, investigated the objective function of EH energy maximization, and a resource allocation problem was formulated to address that goal. Additionally, [58] considered the underlay scheme in CR network and proposed the optimal beamforming design. To address both the SE and EE, a MISO NOMA CR using SWIPT is considered based on a practical non-linear EH model. Robust beamforming design problems are studied under a pair of CSI error models. The related contributions and the motivation of our work are summarized as follows.

5.1.1 Related Work and Motivation

The prior contributions related to this chapter can be divided into two categories based on the EH model adopted, i.e. the linear [55]- [70] and the non-linear EH model [51], [71]- [75]. In the linear EH model, the power harvested increases linearly with the input power, while the EH under the non-linear model exhibits more realistic non-linear characteristics especially at the power-tail.

Linear EH model: In [59], Liu *et al.* analyzed the performance of a cooperative NOMA system relying on SWIPT, which outperformed OMA. Do *et al.* [60] extended [59] and studied the beneficial effect of the user selection scheme on the performance of a cooperative NOMA system using SWIPT. In [61], Yang *et al.* presented a theoretical analysis of two power allocation schemes conceived for a cooperative NOMA system with SWIPT. It was shown that the outage probability achieved under NOMA is lower than that obtained under OMA. Diamantoulakis *et al.* [62] studied the optimal resource allocation design of wireless-powered NOMA systems. The optimal power and time allocation were designed for maximizing the max-min fairness among users. In their following work [63], a joint downlink and uplink scheme was considered in a wireless powered network, followed by comparisons between NOMA and TDMA. The results show that NOMA is more energy efficient in the downlink of SWIPT networks. In order to improve the EE, multiple antennas were applied

in a NOMA system associated with SWIPT, and the transmit beamforming and the power splitting factor were jointly optimized for maximizing the transmit rate of users [64].

The contributions in [59]- [64] investigated conventional wireless NOMA systems, which did not consider the interference between the secondary network and the primary network. Recently, authors of [55], [58], and [68] studied optimal resource allocation problems in CR associated with SWIPT. In [55], an optimal transmit beamforming scheme was proposed in a multi-objective optimization framework. It was shown that there are several tradeoffs in CR-aided SWIPT. Based on the work in [55], the authors proposed a jointly optimal beamforming and power splitting scheme to minimize the transmit power of the base station in multiple-user CR-aided SWIPT [58]. Considering the practical imperfect CSI, Zhou *et al.* [65] studied robust beamforming design problems in MISO CR-aided SWIPT, where the bounded and the gaussian CSI error models were applied. It was shown that the performance achieved under the gaussian CSI error model is better than that obtained under the bounded CSI error model. The work in [65] was then extended to MIMO CR-aided SWIPT in [66] and [67], where the bounded CSI error model was applied in [66] and the gaussian CSI error model was used in [67] and [69]. In contrast to [55], [58]- [67], Zhou *et al.* [68] studied robust resource allocation problems in CR-aided SWIPT under opportunistic spectrum access.

Non-linear EH model: In [51], robust resource allocation schemes were proposed for maximizing the sum transmission rate or the max-min transmission rate of MIMO-assisted wireless powered communication networks, where a practical non-linear EH model is considered. It was shown that a performance gain can be obtained under a practical non-linear EH model over that attained under the linear EH model. In order to maximize the power-efficient and sum-energy harvested by SWIPT systems, Boshkovska *et al.* designed optimal beamforming schemes in [71] and [72]. Recently, under the idealized perfect CSI assumption, the rate-energy region was quantified in MIMO systems relying on SWIPT and the practical non-linear EH model in [73]. In order to improve the security of a SWIPT system, a robust beamforming design problem was studied under a bounded CSI error model in [74]. The investigations in [51], [71]- [74] were performed in the context of conventional

SWIPT systems. Recently, Wang *et al.* [75] extended a range of classic resource allocation problems into a wireless powered CR counterpart. The optimal channel and power allocation scheme were proposed for maximizing the sum transmission rate.

The resource allocation schemes proposed in [59]- [64] investigated a conventional NOMA system with SWIPT. The mutual interference should be considered and the QoS of the PUs should be protected in NOMA CR. Moreover, the resource allocation schemes proposed in [55], [58]- [68] are based on the classic OMA scheme. Thus, these schemes are not applicable to NOMA CR with SWIPT due to the difference between OMA and NOMA. Furthermore, an idealized linear EH model was applied in [55]- [68], which is impractical since the practical power conversion circuit results in a non-linear end-to-end wireless power transfer. Therefore, it is of great importance to design optimal resource allocation schemes for NOMA CR-aided SWIPT based on the practical non-linear EH model.

Although the practical non-linear EH model was applied in [51], [71]- [75], the authors of [51], [71]- [74] considered conventional OMA systems using SWIPT. Moreover, the resource allocation scheme proposed in [68] is based on OMA and cannot be directly introduced in NOMA CR-aided SWIPT. However, at the time of writing, there is a scarcity of investigations on robust resource allocation design for NOMA CR-aided SWIPT under the practical non-linear EH model. Several challenges have to be addressed to design robust resource allocation schemes for NOMA CR-aided SWIPT. For example, the impact of the CSI error and of the residual interference due to the imperfect SIC should be considered, which makes the robust resource allocation problem quite challenging. Thus, we study robust resource allocation problems in NOMA CR-aided SWIPT.

5.1.2 Contributions

Our contribution expands [51] in three major contexts. Firstly, in this chapter, a NOMA MISO CR-aided SWIPT is considered, while a OMA MIMO wireless powered network was used in [51]. Secondly, the work in [51] relies on the bounded CSI error model, while both the bounded and the gaussian CSI error model are applied in our work. Thirdly, part of our work considers the minimum transmit power as the optimization objective, which is

not considered in [51]. Notice that this chapter is also an extension from our conference one [52] which only considered minimizing transmission power under bounded imperfect CSI model. The contributions of our work are hence summarized as follows.

1. A minimum transmission power problem is formulated under both the bounded and the gaussian CSI error models in a NOMA MISO CR network. The robust beamforming weights and the power splitting ratio are jointly designed. The original problem is hard to solve owing to its non-convex nature arising from the non-linear EH model as well as owing to the imperfect CSI. Hence we transform this problem to a convex one. Finally, we prove that the robust beamforming weights can be found and the rank is lower than two under the bounded CSI error model.

2. We also consider another optimization problem, where the objective function is based on maximizing the harvested energy. Similarly, this problem is formulated under the above pair of imperfect CSI error models. The non-linear EH model makes the original problem even harder to solve. Nevertheless, we managed to transform it to an equivalent form and applied a two-loop procedure for solving it. The inner loop solves a convex problem, while the outer loop iteratively adjusts the parameters. Furthermore, to decouple the coupled variables, a one-dimensional search algorithm is proposed as well.

3. Simulation results show the superiority of the proposed scheme over the traditional OMA scheme; the performance gain of NOMA becomes higher when the required data rate at each SU is higher. Moreover, the results also demonstrate that under gaussian CSI error model, the performance is generally better than that under the bounded CSI error model.

5.2 System and Energy Harvesting Models

5.2.1 System Model

We consider a downlink CR system with one cognitive base station (CBS), one primary base station (PBS), N PUs and K SUs. The CBS is equipped with M antennas, while each user and PBS have a single antenna. It is assumed that the SUs are energy-constrained and energy harvest circuits are used. Specifically, the receiver architecture relies on a power splitting design. Once the signal is detected by the receiver, it will be divided into two parts. One part is used for information detection, while the other part for energy harvesting. Similar structures can be found in [59], [64]. To better utilize the radio resources, all UEs are allowed to access the same resource simultaneously. To be specific, the PBS sends messages to all PUs, while the CBS communicates with all SUs simultaneously by applying NOMA principles by controlling the interference from the CBS to PUs below a certain level [46]. Let

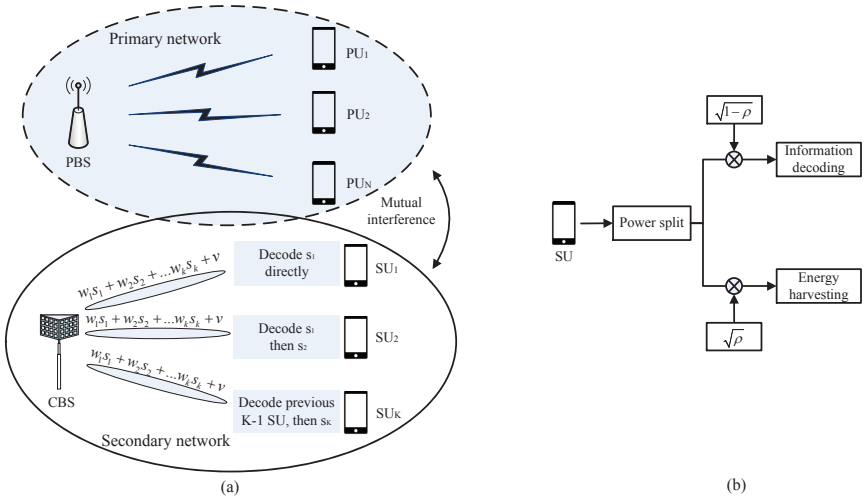

Fig. 5.1: (a) An illustration of the system model. (b) The power splitting architecture of SUs.

us denote the set of SUs and PUs as $\mathcal{K} = \{1, 2, \ldots, K\}$ and $\mathcal{N} = \{1, 2, \ldots, N\}$, respectively. The signal received by the kth SU can be expressed as

$$y_k^S = \mathbf{h}_k^\dagger \mathbf{x} + n_k^S, \ k \in \mathcal{K}, \tag{5.1}$$

where $\mathbf{h}_k \in \mathbb{C}^{M\times 1}$ is the channel gain between the CBS and the kth SU, while n_k^S is

the joint effect of additive white Gaussian noise (AWGN) and interference from the PBS. $n_k^S \sim \mathcal{CN}(0, \sigma_{k,S}^2)$, where $\sigma_{k,S}^2$ is the power. This interference model represents a worst-case scenario [55]. Furthermore, $\mathbf{x}$ is the message transmitted to SUs after precoding. According to the NOMA principle, we have:

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{w}_k s_k + \mathbf{v}, \tag{5.2}$$

where $\mathbf{w}_k \in \mathbb{C}^{M\times 1}$ is the precoding vector for the k-th UE and s_k is the corresponding intended message. Furthermore, $\mathbf{v} \in \mathbb{C}^{M\times 1}$ is the energy vector allowing us to improve the energy harvesting efficiency at the SUs. We assume that s_k is unitary, i.e. $\mathbb{E}[|s_k|^2] = 1$, and $\mathbf{v}$ obeys the complex Gaussian distribution, i.e. $\mathbf{v} \sim \mathcal{CN}(\mathbf{0}, \mathbf{V})$, where $\mathbf{V}$ is the covariance matrix of $\mathbf{v}$.

Likewise, the extra interference arriving from the CBS to the n-th PU is

$$y_n^P = \mathbf{g}_n^\dagger \mathbf{x}, \ n \in \mathcal{N}, \tag{5.3}$$

where $\mathbf{g}_n^\dagger \in \mathbb{C}^{M\times 1}$ is the channel gain between the CBS and the n-th PU [65].

5.2.2 Non-linear EH Model

Most of the existing literature considered an idealized linear energy harvesting model, where the energy collected by the k-th SU is expressed as $E_k^{\text{Linear}} = \eta\ E_k^{\text{In}}$, $E_k^{\text{In}} = \rho\ \left(\mathbf{h}_k^\dagger(\sum_{j=1}^{K} \mathbf{w}_j \mathbf{w}_j^\dagger + \mathbf{V})\mathbf{h}_k + \sigma_{k,S}^2\right)$ is the input power, where ρ is the power splitting factor that controls the amount of received energy allocated to energy harvesting, $0 < \rho < 1$, while η is the energy conversion efficiency factor, $0 < \eta \leq 1$. However, measurements relying on real-world testbeds show that a typical energy harvesting model exhibits a non-linear end-to-end characteristic. To be specific, the harvested energy first grows almost linearly with the increase of the input power, and then saturates when the input power reaches a certain level. Several models have been proposed in the literature and one of the most popular ones

is [51], which is formulated as follows:

$$E_k^{\text{Practical}} = \frac{\Psi_k^{\text{Practical}} - M_k\Omega_k}{1-\Omega_k}, \Omega_k = \frac{1}{1+\exp(a_k b_k)}, \tag{5.4a}$$

$$\Psi_k^{\text{Practical}} = \frac{M_k}{1+\exp\big(-a_k(E_k^{\text{In}} - b_k)\big)}, \tag{5.4b}$$

where $E_k^{\text{Practical}}$ is the actual energy harvested from the circuit. Furthermore, $\Psi_k^{\text{Practical}}$ represents a function of the input power E_k^{In}. Additionally, M_k is the maximum power that a receiver can harvest, while a_k together with b_k characterizes the physical hardware in terms of its circuit sensitivity, limitations, and leakage currents [51].

On the other hand, the signal received in the k-th SU information decoding circuit is

$$y_k^D = \sqrt{1-\rho}(\mathbf{h}_k^\dagger \mathbf{x} + n_k^S) + n_k^D, \tag{5.5}$$

where n_k^D is the AWGN imposed by the information decoding receiver.

5.3 Power Minimization Based Problem Formulation

Since $\mathbf{x}$ is a composite signal consisting of all SUs' messages, SIC is applied at the receiver side to detect the received signal. The detection is carried out in the same order of the channel gains, i.e. the SUs with lower channel gain will be decoded first. A pair of imperfect CSI error models are considered, namely a bounded and a gaussian model. We adopt both of these in this chapter and assume that all SUs have a perfect knowledge of their own CSI.

5.3.1 Bounded CSI Error Model

In this model, we consider a bounded error imposed on the estimated CSI, which can be treated as the worst-case scenario. Specifically, the channels can be modeled as follows.

$$\mathbf{h}_k = \widehat{\mathbf{h}}_k + \Delta\mathbf{h}_k,\ \forall k \in \mathcal{K}, \tag{5.6a}$$

$$\mathbf{\Gamma}_k \triangleq \big\{\Delta\mathbf{h}_k \in \mathbb{C}^{M\times 1} : ||\Delta\mathbf{h}_k||^2 \leq \varphi_k^2\big\}, \tag{5.6b}$$

$$\mathbf{g}_n = \widehat{\mathbf{g}}_n + \Delta\mathbf{g}_n,\ \forall n \in \mathcal{N}, \tag{5.6c}$$

$$\mathbf{\Theta}_n \triangleq \big\{\Delta\mathbf{g}_n \in \mathbb{C}^{M\times 1} : ||\Delta\mathbf{g}_n||^2 \leq \psi_n^2\big\}, \tag{5.6d}$$

where $\widehat{\mathbf{h}}_k$ and $\widehat{\mathbf{g}}_n$ are the estimated channel vectors for $\mathbf{h}_k$ and $\mathbf{g}_n$, respectively, while $\mathbf{\Gamma}_k$ and $\mathbf{\Theta}_n$ define the set of channel variations due to estimation errors. The model defines all the uncertainty regions that are confined by power constraints. Furthermore, we use block Rayleigh fading channels, which remain constant within each block, but change from block to block independently.

NOMA Transmission

Without loss of generality, we sort the estimated channel of SUs in the ascending order, i.e., $||\hat{\mathbf{h}}_1||^2 \leq ||\hat{\mathbf{h}}_2||^2 \leq \ldots\ ||\hat{\mathbf{h}}_K||^2$. According to the SIC principle, SU i can detect and remove SU k's signal, for $1 \leq k < i \leq K$. Thus, when SU i decodes signal s_k, the signals of the previous $(k-1)$ SUs have already been removed from the composite received signal. Due to channel estimation errors, however, these $(k-1)$ signals may not be completely removed, leaving some residual signals as interference. Therefore, the signal at UE i when decoding s_k becomes

$$y_{i,k}^S = \sqrt{1-\rho}\big(\mathbf{h}_i^\dagger\mathbf{w}_k s_k + \sum_{j=1}^{k-1}\Delta\mathbf{h}_i^\dagger\mathbf{w}_j s_j + \sum_{j=k+1}^{K}\mathbf{h}_i^\dagger\mathbf{w}_j s_j + \mathbf{h}_i^\dagger\mathbf{v} + n_k^S\big) + n_k^D. \tag{5.7}$$

Here, the first term is the desired received signal, the second term is the interference due to imperfect channel estimation, and the third term represents the NOMA interference. For notational simplicity, let us denote $\mathbf{W}_k = \mathbf{w}_k\mathbf{w}_k^\dagger$, $\mathbf{V} = \mathbf{v}\mathbf{v}^\dagger$, $S_i^k = \mathbf{h}_i^\dagger\mathbf{W}_k\mathbf{h}_i$, and

$T_i^j = \Delta\mathbf{h}_i^\dagger \mathbf{W}_j \Delta\mathbf{h}_i$. The corresponding signal-to-interference-plus-noise ratio (SINR) for the i-th SU after the SIC applied at the receiver is given by:

$$\text{SINR}_i^k = \frac{S_i^k}{\sum_{j=1}^{k-1} T_i^j + \sum_{j=k+1}^{K} S_i^j + \mathbf{h}_i^\dagger \mathbf{V} \mathbf{h}_i + \sigma_{k,S}^2 + \frac{\sigma_D^2}{(1-\rho)}}.$$

Since the signal s_k can be detected at every SU i, as long as $k < i$, there will be a set of SINRs for signal s_k. For CBS, the maximum data rate for SU k should be $R_k = \log_2(1 + \min_{k \le i \le K} \text{SINR}_i^k)$. Moreover, the channel estimation error should be considered. The worst-case data rate for SU k becomes

$$R_k = \log_2 \big(1 + \min_{\Delta\mathbf{h}_i \in \mathbf{\Gamma}_i} \{ \min_{k \le i \le K} \text{SINR}_i^k \}\big). \tag{5.8}$$

Problem Formulation

In this sub-section, we seek to find the precoding vectors $\mathbf{w}_k,\ k \in \mathcal{K}$, the energy vector $\mathbf{v}$, and the power split ratio ρ, which altogether achieve a satisfactory QoS for all users, and at the same time, they can harvest part of the energy for their future usage. Thus, the problem can be formulated as follows:

$$\mathbf{P}_1: \quad \min_{\mathbf{W}_k \in \mathbb{C}^{M \times M}, \mathbf{V} \in \mathbb{C}^{M \times M}, \rho} \ \text{Tr}\Big(\sum_{k=1}^{K} \mathbf{W}_k + \mathbf{V}\Big) \tag{5.9a}$$

$$\text{s.t.} C1: \quad R_k \ge R_{k,\min} \tag{5.9b}$$

$$C2: \quad E_k^{\text{Practical}} \ge P_{k,s},\ \forall \Delta\mathbf{h}_k \in \mathbf{\Gamma}_k,\ \forall k \in \mathcal{K}, \tag{5.9c}$$

$$C3: \quad \mathbf{g}_n^\dagger \Big(\sum_{j=1}^{K} \mathbf{W}_j + \mathbf{V}\Big) \mathbf{g}_n \le P_{n,p},\ \forall \Delta\mathbf{g}_n \in \mathbf{\Theta}_n, \tag{5.9d}$$

$$C4: \quad \text{Tr}\Big(\sum_{k=1}^{K} \mathbf{W}_k + \mathbf{V}\Big) \le P_B, \tag{5.9e}$$

$$C5: \quad 0 < \rho < 1, \tag{5.9f}$$

$$C6: \quad \mathbf{V} \succeq \mathbf{0}, \mathbf{W}_k \succeq \mathbf{0}, \tag{5.9g}$$

$$C7: \quad \text{Rank}(\mathbf{W}_k) = 1,\ \forall k \in \mathcal{K}. \tag{5.9h}$$

Our goal is to minimize the total transmitted power. The constraint $C1$ ensures that SU k does attain the predefined minimum data rate; $C2$ allows each SU to harvest the amount of energy that at least compensates the static power dissipation $P_{k,s}$; $C3$ is the interference limit for the n-th PU; $C4$ represents the maximum transmit power constraint of the BS; in $C5$, the power split factor should be in the range of $(0, 1)$. The optimization problem $\mathbf{P}_1$ is hard to solve due to its non-convexity constraints $C1$ and $C2$. Moreover, the realistic imperfect CSI imposes another challenge on the original problem. In the following, we transform the variables.

Let us introduce $\gamma_{k,\min} \triangleq (2^{R_{k,\min}} - 1)$. Then $C1$ in (5.9b) becomes

$$\min_{\Delta\mathbf{h}_i \in \mathbf{\Gamma}_i} \frac{S_i^k}{\sum_{j=1}^{k-1} T_i^j + \sum_{j=k+1}^{K} S_i^j + \mathbf{h}_i^\dagger \mathbf{V} \mathbf{h}_i + \sigma_{k,S}^2 + \frac{\sigma_D^2}{(1-\rho)}} \geq \gamma_{k,\min}, \tag{5.10}$$

where $i = \{k, k+1, \ldots, K\}, \forall k \in \mathcal{K}$. For the notational simplicity, we denote the above constraint as $\Xi_{i,k}$. Thus, $\mathbf{P}_1$ becomes

$$\mathbf{P}_2: \quad \min_{\mathbf{W}_k \in \mathbb{C}^{M\times M}, \mathbf{V} \in \mathbb{C}^{M\times M}, \rho} \ \mathrm{Tr}\Big(\sum_{k=1}^{K} \mathbf{W}_k + \mathbf{V}\Big) \tag{5.11a}$$

$$\text{s.t. } C1: \quad \Xi_{i,k} \tag{5.11b}$$

$$C2: \quad E_k^{\text{Practical}} \geq P_{k,s}, \ \forall \Delta\mathbf{h}_k \in \mathbf{\Gamma}_k, \ \forall k \in \mathcal{K}, \tag{5.11c}$$

$$C3: \quad \mathbf{g}_n^\dagger \Big(\sum_{j=1}^{K} \mathbf{W}_j + \mathbf{V}\Big) \mathbf{g}_n \leq P_{n,p}, \ \forall \Delta\mathbf{g}_n \in \mathbf{\Theta}_n, \tag{5.11d}$$

$$C4: \quad \mathrm{Tr}\Big(\sum_{k=1}^{K} \mathbf{W}_k + \mathbf{V}\Big) \leq P_B, \tag{5.11e}$$

$$C5: \quad 0 < \rho < 1, \tag{5.11f}$$

$$C6: \quad \mathbf{V} \succeq \mathbf{0}, \mathbf{W}_k \succeq \mathbf{0}, \tag{5.11g}$$

$$C7: \quad \mathrm{Rank}(\mathbf{W}_k) = 1, \ \forall k \in \mathcal{K}. \tag{5.11h}$$

Here, $C6$ comes from the fact that both $\mathbf{V}$ and $\mathbf{W}_k$ are positive semi-definite matrices. The extra constraint that the rank of $\mathbf{W}_k$ should be 1 is also non-convex. In what follows, we first reformulate $C1$ in (5.11b) according to the $\mathcal{S}$-Procedure of [76].

Lemma 2. *$C1$ in (5.11b) can be reformulated as*

$$\begin{bmatrix} \alpha_{i,k}\mathbf{I} + \mathbf{C}_k - \gamma_{k,min}\sum_{j=1}^{k-1}\mathbf{W}_j & \mathbf{C}_k\widehat{\mathbf{h}}_i \\ \widehat{\mathbf{h}}_i^{\dagger}\mathbf{C}_k & -\alpha_{i,k}\varphi_k^2 + \Phi_k \end{bmatrix} \succeq \mathbf{0}, \tag{5.12}$$

$\forall k \in \mathcal{K},\ i = \{k, k+1, \ldots, K\}$, where $\mathbf{C}_k = \mathbf{W}_k - \gamma_{k,min}(\sum_{j=k+1}^{K}\mathbf{W}_j + \mathbf{V})$ and $\Phi_k = \widehat{\mathbf{h}}_i^{\dagger}\mathbf{C}_k\widehat{\mathbf{h}}_i - \gamma_{k,min}\big(\sigma_{k,S}^2 + \frac{\sigma_D^2}{(1-\rho)}\big)$, and $\alpha_{i,k}$ is a slack variable conditioned on $\alpha_{i,k} \geq 0$.

Proof. Given $\mathbf{h}_i = \widehat{\mathbf{h}}_i + \Delta\mathbf{h}_i$ and (5.10), we have

$$\begin{aligned} &\Delta\mathbf{h}_i^{\dagger}\big(\gamma_{k,\min}(\sum_{j\neq k}\mathbf{W}_j + \mathbf{V}) - \mathbf{W}_k\big)\Delta\mathbf{h}_i + 2\ \mathrm{Re}\big\{\widehat{\mathbf{h}}_i^{\dagger}\big(\gamma_{k,\min}(\sum_{j=k+1}^{K}\mathbf{W}_j + \mathbf{V}) - \mathbf{W}_k\big)\Delta\mathbf{h}_i\big\} \\ &+ 2\widehat{\mathbf{h}}_i^{\dagger}\big(\gamma_{k,\min}(\sum_{j=k+1}^{K}\mathbf{W}_j + \mathbf{V}) - \mathbf{W}_k\big)\widehat{\mathbf{h}}_i + 2\gamma_{k,\min}\big(\sigma_{k,S}^2 + \frac{\sigma_D^2}{(1-\rho)}\big) \leq 0. \end{aligned} \tag{5.13}$$

From the fact that $\Delta\mathbf{h}_i^{\dagger}\Delta\mathbf{h}_i - \varphi_k^2 \leq 0$ and according to the $\mathcal{S}$-Procedure, the lemma is proved. □

Similarly, $C3$ in (5.11d) can be transformed into

$$\begin{bmatrix} \beta_n\mathbf{I} - \boldsymbol{\Sigma} & -\boldsymbol{\Sigma}\widehat{\mathbf{g}}_n \\ -\widehat{\mathbf{g}}_n^{\dagger}\boldsymbol{\Sigma} & -\beta_n\psi_n^2 - \widehat{\mathbf{g}}_n^{\dagger}\boldsymbol{\Sigma}\widehat{\mathbf{g}}_n + P_{n,p} \end{bmatrix} \succeq \mathbf{0},\ \forall n \in \mathcal{N}, \tag{5.14}$$

where $\boldsymbol{\Sigma} = \sum_{j=1}^{K}\mathbf{W}_j + \mathbf{V}$, and $\beta_n \geq 0$ is also a slack variable.

Next, we apply similar manipulations to (5.11c), which becomes

$$\min_{\Delta\mathbf{h}_k \in \boldsymbol{\Gamma}_k} \rho\ \big(\mathbf{h}_k^{\dagger}\boldsymbol{\Sigma}\mathbf{h}_k + \sigma_{k,S}^2\big) \quad \geq \quad D_k, \tag{5.15}$$

where $D_k = -\ln\big(\frac{1}{P_{k,s}(1-\Omega_k)/M_k + \Omega_k} - 1\big)/a_k + b_k$ is a constant. This condition holds, provided that $a_k > 0$, which is always true in real systems.

Then, applying the $\mathcal{S}$-Procedure to (5.15), we have the following

$$\begin{bmatrix} \theta_k \mathbf{I} + \mathbf{\Sigma} & \mathbf{\Sigma}\widehat{\mathbf{h}}_k \\ \widehat{\mathbf{h}}_k^\dagger \mathbf{\Sigma} & -\theta_k \varphi_k^2 + \widehat{\mathbf{h}}_k^\dagger \mathbf{\Sigma} \widehat{\mathbf{h}}_k + \sigma_{k,S}^2 - \frac{D_k}{\rho} \end{bmatrix} \succeq \mathbf{0}, \tag{5.16}$$

$\forall k \in \mathcal{K}$, where $\theta_k \geq 0$.

Therefore, $\mathbf{P}_2$ becomes

$$\mathbf{P}_3: \quad \min_{\mathbf{W}_k, \mathbf{V}, \rho, \{\alpha_{i,k}\}, \{\beta_n\}, \{\theta_k\}} \ \mathrm{Tr}\Big(\sum_{k=1}^{K} \mathbf{W}_k + \mathbf{V}\Big) \tag{5.17a}$$

$$\text{s.t.} \quad (5.12), (5.14), (5.16), (5.11e), (5.11f), (5.11g), \tag{5.17b}$$

$$\alpha_{i,k}, \beta_n, \theta_k \geq 0, \tag{5.17c}$$

$$\forall k \in \mathcal{K}, \ i = \{k, k+1, \ldots, K\}, \ \forall n \in \mathcal{K}.$$

Observe that we drop (5.11h), since it is not a convex term. This relaxation is commonly referred to as the semi-definite relaxation (SDR) technique. For the specific problem in $\mathbf{P}_2$, the following theorem proves that the optimal $\mathbf{W}_k$ has a limited rank.

Theorem 6. *If $\mathbf{P}_2$ is feasible, the rank of $\mathbf{W}_k, k \in \mathcal{K}$ is always less than or equal to 2.*

Proof. See Appendix. □

The transformed problem $\mathbf{P}_3$ is not convex because of the coupling variables ρ in (5.16) and $(1-\rho)$ in the denominator of (5.12). To be able to take advantage of the *CVX* software package, we introduce a pair of auxiliary variables. Specifically, let $p = \frac{1}{1-\rho}$ and $q = \frac{1}{\rho}$. In this way, (5.12), (5.14), and (5.16) become convex terms. Then, we have additional constraints for p and q:

$$p \geq \frac{1}{1-\rho} \quad \text{and} \quad q \geq \frac{1}{\rho}. \tag{5.18}$$

It may be readily verified that this transformation does not change the optimal solution of $\mathbf{P}_3$.

5.3.2 Matrix Decomposition

Now we proceed to find the solution of the problem $\mathbf{P}_2$, after which there is one more step to get the original solution for $\mathbf{w}_k$. If $\mathbf{W}_k$ yields rank 1, we can simply write $\mathbf{W}_k^\star = \mathbf{w}_k^\star \mathbf{w}_k^{\star\dagger}$. Otherwise, if $\text{Rank}(\mathbf{W}_k^\star) = 2$, we have several optional approaches to extract $\mathbf{w}_k^\star$. To name a few, we list two methodologies here.

1. *Eigen-decomposition.* Let us denote two eigenvalues of $\mathbf{W}_k^\star$ by λ_1 and λ_2, where $\lambda_1 > \lambda_2 \geq 0$. Clearly, $\mathbf{W}_k^\star = \lambda_1 \mathbf{w}_{1k}\mathbf{w}_{1k}^\dagger + \lambda_2 \mathbf{w}_{2k}\mathbf{w}_{2k}^\dagger$, $\mathbf{w}_{ik}, i = \{1, 2\}$ are the corresponding eigenvectors. To get the rank 1 approximation from a rank 2 matrix, we can let the solution of the original problem be $\widehat{\mathbf{w}}_k = \sqrt{\lambda_1}\mathbf{w}_{1k}\mathbf{w}_{1k}^\dagger$, provided it is feasible.

2. *Randomization technique.* Similar to eigen-decomposition, we first decompose $\mathbf{W}_k^\star$ according to $\mathbf{W}_k^\star = \mathbf{U}_k \mathbf{T}_k \mathbf{U}_k^\dagger$. Then, we let $\widehat{\mathbf{w}}_k = \mathbf{U}_k \mathbf{T}_k^{1/2} \mathbf{e}_k$, where the m-th element of $\mathbf{e}_k$ is $[\mathbf{e}_k]_m = e^{j\theta_{k,m}}$ and $\theta_{k,m}$ obeys an independent and uniform distribution within $[0, 2\pi)$.

The above two methods are essentially the same. If we want to get a more precise result, another scaling factor can be added. Specifically, let us define c_k as the scaling factor yet to be determined. Certainly, the problem can be transformed in terms of $\mathbf{W}_k$ and c_k, once we get the optimal value, we can apply either one of the above methods to get a better result. Another point worth noting here is that when the rank of $\mathbf{W}_k$ is 2, there only exists the approximation result of $\mathbf{w}_k^\star$, and this approximation always provides an upper bound.

5.3.3 Gaussian CSI Error Model

In Section III-*A*, we introduced a bounded channel model, which defines a confined region for the channel variations, which provides a worst-case estimation. Another commonly used more realistic estimation model assumes that the channel estimation error obeys the Gaussian distribution [65] [70] [75], which is formulated as follows:

$$\mathbf{h}_k = \widehat{\mathbf{h}}_k + \Delta\mathbf{h}_k, \ \Delta\mathbf{h}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{H}_k), \ \forall k \in \mathcal{K}, \tag{5.19a}$$

$$\mathbf{g}_n = \widehat{\mathbf{g}}_n + \Delta\mathbf{g}_n, \ \Delta\mathbf{g}_n \sim \mathcal{CN}(\mathbf{0}, \mathbf{G}_n), \ \forall n \in \mathcal{N}, \tag{5.19b}$$

where $\Delta\mathbf{h}_k$ and $\Delta\mathbf{g}_n$ are the channel estimation error vectors, while $\widehat{\mathbf{h}}_k$ and $\widehat{\mathbf{g}}_n$ are the channel vectors estimated at the BS side. Furthermore, $\mathbf{H}_k$ and $\mathbf{G}_n$ are the covariance matrices of the estimation error vectors.

Even though we apply different channel models, the residual interference due to imperfect CSI estimation affects the message detection similarly to the bounded error model. Thus the achievable data rate expression of SU k remains the same except that $\Delta\mathbf{h}_k$ is in a new set. In contrast to the existing NOMA contributions on imperfect CSI [33], in this chapter we use the above-mentioned gaussian estimation error model to form an optimization problem as follows:

$$\mathbf{P}_4: \quad \min_{\mathbf{W}_k\in\mathbb{C}^{M\times M},\mathbf{V}\in\mathbb{C}^{M\times M},\rho} \ \mathrm{Tr}(\sum_{k=1}^{K}\mathbf{W}_k+\mathbf{V}) \tag{5.20a}$$

$$\text{s.t. } C1: \quad \Pr\{R_k \geq R_{k,\min}\} \geq 1-\xi_k, \ \forall k\in\mathcal{K}, \tag{5.20b}$$

$$C2: \quad \Pr\{E_k^{\text{Practical}} \geq P_{k,s}\} \geq 1-\xi_{k,s}, \tag{5.20c}$$

$$\forall\Delta\mathbf{h}_k \sim \mathcal{CN}(\mathbf{0},\mathbf{H}_k), \ \forall k\in\mathcal{K},$$

$$C3: \quad \Pr\left\{\mathbf{g}_n^{\dagger}\mathbf{\Sigma}\mathbf{g}_n \leq P_{n,p}\right\} \geq 1-\xi_{n,p}, \tag{5.20d}$$

$$\forall\Delta\mathbf{g}_n \sim \mathcal{CN}(\mathbf{0},\mathbf{G}_n), \forall n\in\mathcal{N},$$

$$C4: \quad (5.11e)-(5.11h). \tag{5.20e}$$

Here, we assume that the probability of having a rate of R_k is higher than $R_{k,\min}$, which is a predefined value, and we use the threshold ξ_k to control the probability. Likewise, $\xi_{k,s}$ and $\xi_{n,p}$, where $k\in\mathcal{K}$ and $n\in\mathcal{N}$, are used for controlling the outage probability of harvested energy of the kth SU and the interference experienced by the n-th PU, respectively. $\mathbf{P}_4$ is hard to solve owing to its non-convexity, together with constraints $C1-C3$, which involve probability and uncertainty. Inspired by [65], we solve the resulted optimization problem with the aid of approximations by applying Bernstein-type inequalities [77].

Bernstein-type Inequality I [77]

Let $f(\mathbf{z}) = \mathbf{z}^\dagger \mathbf{A}\mathbf{z} + 2\mathrm{Re}\{\mathbf{z}^\dagger \mathbf{b}\} + c$, where $\mathbf{A} \in \mathbb{H}^N$, $\mathbf{b} \in \mathbb{C}^{N\times 1}$, $c \in \mathbb{R}$, and $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$. For any $\xi \in (0, 1]$, an approximate and convex form of

$$\Pr\{f(\mathbf{z}) \geq 0\} \geq 1 - \xi \tag{5.21}$$

can be written as

$$\mathrm{Tr}(\mathbf{A}) - \sqrt{-2\ln(\xi)}\upsilon_1 + \ln(\xi)\upsilon_2 + c \geq 0, \tag{5.22a}$$

$$\left|\left| \begin{bmatrix} \mathrm{vec}(\mathbf{A}) \\ \sqrt{2}\mathbf{b} \end{bmatrix} \right|\right| \leq \upsilon_1, \tag{5.22b}$$

$$\upsilon_2 \mathbf{I} + \mathbf{A} \succeq \mathbf{0}, \upsilon_2 \geq 0. \tag{5.22c}$$

Here, υ_1 and υ_2 are slack variables.

In order to use the above Lemma, we have to transform $\Delta\mathbf{h}_i$ to a standard complex Gaussian vector. Let $\Delta\mathbf{h}_i = \mathbf{H}_i^{1/2}\tilde{\mathbf{h}}_i$, where $\tilde{\mathbf{h}}_i \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$. Substituting it into (5.10), the convex approximation becomes

$$\mathrm{Tr}\big(\mathbf{H}_i^{1/2}(\mathbf{C}_k - \gamma_{k,\min}\sum_{j=1}^{k-1}\mathbf{W}_j)\mathbf{H}_i^{1/2}\big) - \sqrt{-2\ln(\xi_k)}\upsilon_{1i,k} + \ln(\xi_k)\upsilon_{2i,k} + c_{i,k} \geq 0, \tag{5.23a}$$

$$c_{i,k} = \widehat{\mathbf{h}}_i^\dagger \mathbf{C}_k \widehat{\mathbf{h}}_i - r_{k,\min}\big(\sigma_{k,S}^2 + \frac{\sigma_D^2}{1-\rho}\big), \tag{5.23b}$$

$$\left|\left| \begin{bmatrix} \mathrm{vec}\big(\mathbf{H}_i^{1/2}(\mathbf{C}_k - \gamma_{k,\min}\sum_{j=1}^{k-1}\mathbf{W}_j)\mathbf{H}_i^{1/2}\big) \\ \sqrt{2}\mathbf{H}_i^{1/2}\mathbf{C}_k\widehat{\mathbf{h}}_i \end{bmatrix} \right|\right| \leq \upsilon_{1i,k}, \tag{5.23c}$$

$$\upsilon_{2i,k}\mathbf{I} + \big(\mathbf{H}_i^{1/2}(\mathbf{C}_k - \gamma_{k,\min}\sum_{j=1}^{k-1}\mathbf{W}_j)\mathbf{H}_i^{1/2}\big) \succeq \mathbf{0}, \tag{5.23d}$$

$$\upsilon_{2i,k} \geq 0, \forall k \in \mathcal{K}, i = \{k, \ldots, K\},$$

where $\upsilon_{1i,k}$ and $\upsilon_{2i,k}$ are slack variables.

For (5.20d), we use a simple transformation similar as that in (5.15), which leads to:

$$\Pr\{\rho(\mathbf{h}_k^\dagger \mathbf{\Sigma} \mathbf{h}_k + \sigma_{k,S}^2) \geq D_k\} \geq 1 - \xi_{k,s}. \tag{5.24}$$

Furthermore, by applying the inequalities in (5.22), (5.24) can be expressed as

$$\mathrm{Tr}(\mathbf{H}_k^{1/2} \mathbf{\Sigma} \mathbf{H}_k^{1/2}) - \sqrt{-2\ln(\xi_{k,s})}\upsilon_{1k,s} + \ln(\xi_{k,s})\upsilon_{2k,s} + c_{k,s} \geq 0, \tag{5.25a}$$

$$c_{k,s} = \widehat{\mathbf{h}}_k^\dagger \mathbf{\Sigma} \widehat{\mathbf{h}}_k + \sigma_{k,S}^2 - \frac{D_k}{\rho}, \tag{5.25b}$$

$$\left\| \begin{bmatrix} \mathrm{vec}(\mathbf{H}_k^{1/2} \mathbf{\Sigma} \mathbf{H}_k^{1/2}) \\ \sqrt{2}\mathbf{H}_k^{1/2} \mathbf{\Sigma} \widehat{\mathbf{h}}_k \end{bmatrix} \right\| \leq \upsilon_{1k,s}, \tag{5.25c}$$

$$\upsilon_{2k,s}\mathbf{I} + (\mathbf{H}_k^{1/2} \mathbf{\Sigma} \mathbf{H}_k^{1/2}) \succeq \mathbf{0}, \upsilon_{2k,s} \geq 0, \ \forall k \in \mathcal{K}, \tag{5.25d}$$

where $\upsilon_{1k,s}$ and $\upsilon_{2k,s}$, $k \in \mathcal{K}$, are slack variables.

Bernstein-type Inequality II [78]

Let $f(\mathbf{z}) = \mathbf{z}^\dagger \mathbf{A} \mathbf{z} + 2\mathrm{Re}\{\mathbf{z}^\dagger \mathbf{b}\} + c$, where $\mathbf{A} \in \mathbb{H}^N$, $\mathbf{b} \in \mathbb{C}^{N \times 1}$, $c \in \mathbb{R}$, and $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$. For any $\xi \in (0, 1]$, an approximate and convex form for

$$\Pr\{f(\mathbf{z}) \leq 0\} \geq 1 - \xi \tag{5.26}$$

can be written as

$$\mathrm{Tr}(\mathbf{A}) + \sqrt{-2\ln(\xi)}\upsilon_1 - \ln(\xi)\upsilon_2 + c \geq 0, \tag{5.27a}$$

$$\left\| \begin{bmatrix} \mathrm{vec}(\mathbf{A}) \\ \sqrt{2}\mathbf{b} \end{bmatrix} \right\| \leq \upsilon_1, \tag{5.27b}$$

$$\upsilon_2 \mathbf{I} - \mathbf{A} \succeq \mathbf{0}, \upsilon_2 \geq 0, \tag{5.27c}$$

where υ_1 and υ_2 are slack variables.

We apply Bernstein-type Inequality II to (5.20e), and let $\Delta \mathbf{g}_n = \mathbf{G}_n^{1/2} \tilde{\mathbf{g}}_n$, where $\tilde{\mathbf{g}}_n \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$ is a standard Gaussian vector. We can have the following convex-form

approximation.

$$\mathrm{Tr}(\mathbf{G}_n^{1/2}\mathbf{\Sigma}\mathbf{G}_n^{1/2}) + \sqrt{-2\ln(\xi_{n,p})}\upsilon_{1,n} - \ln(\xi_{n,p})\upsilon_{2,n} + c_n \geq 0, \tag{5.28a}$$

$$c_n = \widehat{\mathbf{g}}_n^{\dagger}\mathbf{\Sigma}\widehat{\mathbf{g}}_n - P_{n,p}, \tag{5.28b}$$

$$\left\| \begin{bmatrix} \mathrm{vec}(\mathbf{G}_n^{1/2}\mathbf{\Sigma}\mathbf{G}_n^{1/2}) \\ \sqrt{2}\mathbf{G}_n^{1/2}\mathbf{\Sigma}\widehat{\mathbf{g}}_n \end{bmatrix} \right\| \leq \upsilon_{1,n}, \tag{5.28c}$$

$$\upsilon_{2,n}\mathbf{I} - \mathbf{G}_n^{1/2}\mathbf{\Sigma}\mathbf{G}_n^{1/2} \succeq \mathbf{0}, \upsilon_{2,n} \geq 0, \forall n \in \mathcal{N}, \tag{5.28d}$$

where $\upsilon_{1,n}$ and $\upsilon_{2,n}$ are slack variables.

Lastly, we relax $\mathbf{P}_4$ by dropping the constraint that $\mathbf{W}_k$ should have rank 1 for now, since it is not a convex one. The relaxed version of the problem is

$$\mathbf{P}_5: \quad \min_{\substack{\mathbf{W}_k,\mathbf{V},\rho,\{\upsilon_{1i,k}\},\{\upsilon_{2i,k}\},\\ \{\upsilon_{1k,s}\},\{\upsilon_{2k,s}\},\{\upsilon_{1,n}\},\{\upsilon_{2,n}\}}} \mathrm{Tr}\Big(\sum_{k=1}^{K}\mathbf{W}_k + \mathbf{V}\Big) \tag{5.29a}$$

$$\text{s.t.} \quad (5.23), (5.25), (5.28), (5.11e), (5.11f), (5.11g). \tag{5.29b}$$

Likewise, the coupling variables in (5.23b) and (5.25b) make $\mathbf{P}_5$ a non-convex problem. Thus we can still use the transformation in (5.18), which converts $\mathbf{P}_5$ into an equivalent optimization problem.

5.4 Maximum Harvested Energy Problem Formulation

In contrast to Sections III, where the minimum transmission power problem is considered, in the following we consider the optimization problem of maximizing the total harvested energy. This problem has important real-world applications, since most of the consumer electronics products are battery-driven and thus their energy efficiency is critical. In this section, we first formulate the problem, then we transform it in a convex way so that an existing software package can solve it efficiently. A one-dimensional search algorithm will be used. Furthermore, we also consider our previous pair of channel models.

5.4.1 Bounded CSI Error Model

Upon considering the imperfect CSI model used in (5.6), the maximum total harvested energy of all SUs can be formulated as follows:

$$\mathbf{P}_6: \quad \max_{\substack{\mathbf{W}_k\in\mathbb{C}^{M\times M},\mathbf{V}\in\mathbb{C}^{M\times M},\\ \rho,\{\alpha_{i,k}\},\{\beta_n\},\{\theta_k\}}} \sum_{k=1}^{K} E_k^{\text{Practical}} \tag{5.30a}$$

$$\text{s.t.} \quad (5.12), (5.14), (5.11e), (5.11f), (5.11g), (5.11h), \tag{5.30b}$$

$$\alpha_{i,k}, \beta_n, \theta_k \geq 0, \ \forall k \in \mathcal{K}, \ i = \{k, k+1, \ldots, K\}, \ \forall n \in \mathcal{K}. \tag{5.30c}$$

The rank operation is not convex, thus we drop the constraint (5.11h) first, as previously in $\mathbf{P}_3$. Additionally, the objective function relies on a realistic non-linear energy harvesting model, and it is not convex either. Essentially, it is a sum-of-ratio problem, and its global optimization is possible by applying the following transformations:

$$\max_{\substack{\mathbf{W}_k\in\mathbb{C}^{M\times M},\mathbf{V}\in\mathbb{C}^{M\times M}\\ \rho,\{\alpha_{i,k}\},\{\beta_n\},\{\theta_k\},\{\tau_k\}}} \sum_{k=1}^{K} \frac{M_k}{1+\exp\big(-a_k(\tau_k-b_k)\big)} \tag{5.31a}$$

$$E_k^{\text{In}} \geq \tau_k, \ \forall \Delta\mathbf{h}_k. \ \forall k \in \mathcal{K}. \tag{5.31b}$$

After applying the $\mathcal{S}$-Procedure of [76] to (5.31b), it becomes

$$\begin{bmatrix} \theta_k\mathbf{I}+\boldsymbol{\Sigma} & \boldsymbol{\Sigma}\widehat{\mathbf{h}}_k \\ \widehat{\mathbf{h}}_k^{\dagger}\boldsymbol{\Sigma} & -\theta_k\varphi_k^2+\widehat{\mathbf{h}}_k^{\dagger}\boldsymbol{\Sigma}\widehat{\mathbf{h}}_k+\sigma_{k,S}^2-\frac{\tau_k}{\rho} \end{bmatrix} \succeq \mathbf{0}, \tag{5.32}$$

$\forall k \in \mathcal{K}$. Furthermore, according to [71], [79], if $\mathbf{P}_6$ has the optimal solutions $\mathbf{W}_k^{\star}$ and $\mathbf{V}^{\star}$, there exist two sets of vectors $\boldsymbol{\mu} = \{\mu_1, \mu_2, \ldots, \mu_K\}$ and $\boldsymbol{\epsilon} = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_K\}$ such that the solutions are also optimal for the following equivalent parametric optimization problem:

$$\mathbf{P}_7 \quad \max_{\substack{\mathbf{W}_k\in\mathbb{C}^{M\times M}\\ \mathbf{V}\in\mathbb{C}^{M\times M}\\ \rho,\{\alpha_{i,k}\},\{\beta_n\},\{\theta_k\},\{\tau_k\}}} \sum_{k=1}^{K} \mu_k\big\{M_k-\epsilon_k\big(1+\exp(-a_k(\tau_k-b_k))\big)\big\}. \tag{5.33}$$

The optimal solutions and the vectors should satisfy

$$\epsilon_k\big(1+\exp(-a_k(\tau_k^\star - b_k))\big) - M_k = 0, \tag{5.34a}$$

$$\mu_k\big(1+\exp(-a_k(\tau_k^\star - b_k))\big) - 1 = 0, \forall k \in \mathcal{K}, \tag{5.34b}$$

where $E_k^{\text{In},\star} = \rho^\star \left(\mathbf{h}_k^\dagger(\sum_{j=1}^K \mathbf{W}_j^\star + \mathbf{V}^\star)\mathbf{h}_k + \sigma_{k,S}^2\right) \geq \tau_k^\star$.

Now, the objective function has the *log-concave* form and it can be solved given the sets $\boldsymbol{\mu}$ and $\boldsymbol{\epsilon}$. The iterative update of the vector sets can be carried out in the following way. Let us define the function $\mathcal{F}(\boldsymbol{\mu},\boldsymbol{\epsilon}) = \big[\epsilon_k\big(1+\exp(-a_k(\tau_k^\star - b_k))\big) - M_k, \ldots, \mu_k\big(1+\exp(-a_k(\tau_k^\star - b_k))\big) - 1\big]$, $\forall k \in \mathcal{K}$. The next set of values of $\boldsymbol{\mu}$ and $\boldsymbol{\epsilon}$ can be updated by solving $\mathcal{F}(\boldsymbol{\mu},\boldsymbol{\epsilon}) = \mathbf{0}$. Specifically, in the q-th iteration, we update them as:

$$\boldsymbol{\mu}^{q+1} = \boldsymbol{\mu}^q + \varpi^q \mathbf{p}^q, \;\; \boldsymbol{\epsilon}^{q+1} = \boldsymbol{\epsilon}^q + \varpi^q \mathbf{p}^q, \tag{5.35}$$

where $\mathbf{p}^q = [\mathcal{F}'(\boldsymbol{\mu},\boldsymbol{\epsilon})]\mathcal{F}(\boldsymbol{\mu},\boldsymbol{\epsilon})$, $\mathcal{F}'(\boldsymbol{\mu},\boldsymbol{\epsilon})$ is the Jacobian matrix of $\mathcal{F}(\boldsymbol{\mu},\boldsymbol{\epsilon})$, ϖ^q is the largest ϖ^l that satisfies $||\mathcal{F}(\boldsymbol{\mu}^q + \varpi^l\mathbf{p}^q, \boldsymbol{\epsilon}^q + \varpi^l\mathbf{p}^q)|| \leq (1 - t\varpi^l)||\mathcal{F}(\boldsymbol{\mu},\boldsymbol{\epsilon})||$, $l = 1, 2, \ldots$, $0 < \varpi^l < 1$, and $0 < t < 1$ [71] [79].

A two-loop algorithm is proposed for solving the problem. The outer loop gives $\boldsymbol{\mu}$ and $\boldsymbol{\epsilon}$ as the inputs of the inner loop, while the inner loop finds $\mathbf{W}_k^\star$ and $\mathbf{V}^\star$. Observe that in (5.32), there is a coupling variable $\frac{\tau_k}{\rho}$, which is convex with a given ρ. Therefore, in the inner loop, we have to perform a one-dimensional search for ρ as well. The detailed algorithm is formulated in Algorithm 1.

5.4.2 Gaussian CSI Error Model

In this section, we formulate the maximum harvested energy under the gaussian CSI error model formulated is as follows:

$$\mathbf{P}_8: \quad \max_{\mathbf{W}_k \in \mathbb{C}^{M\times M}, \mathbf{V} \in \mathbb{C}^{M\times M}, \rho} \; \sum_{k=1}^K E_k^{\text{Practical}} \tag{5.36a}$$

$$\text{s.t.} \quad (5.20b), (5.20e), (5.11e), (5.11f), (5.11g), (5.11h). \tag{5.36b}$$

Algorithm 2 Robust Precoding Design for EH Maximization Problem

Input: Minimum required data rate R_k of SU k, noise power $\sigma_{k,S}^2$ and σ_D^2, channel uncertainty φ_k^2 and ψ_n^2, maximum allowed interference power $P_{n,p}$ for PU n, maximum BS transmitted power P_B, and randomly generated estimated channel $\widehat{\mathbf{h}}_k$ and $\widehat{\mathbf{g}}_n$.
Initialisation: Iteration number $q = 0, p = 1$, initial value of ρ as ρ_{start}, step s, end value ρ_{end}, $\boldsymbol{\mu}^0$, and $\boldsymbol{\epsilon}^0$, loop stop criteria m_{th}.
One-dimensional Search:
for $\rho = \rho_{\text{start}} : s : \rho_{\text{end}}$ **do**
  repeat: {Outer Loop}
    Solve for the optimization problem $\mathbf{P}_7$: {Inner Loop}
    if ($\mathbf{P}_7$ is feasible) **then**
      Obtain $\mathbf{W}_k^q$ and $\mathbf{V}^q$.
    else
      Break from the outer loop.
    end if
    Update $\boldsymbol{\mu}^{q+1}$ and $\boldsymbol{\epsilon}^{q+1}$ according to (5.35), then let $q = q + 1$.
  until $\left|\mu_k^{q+1}\left\{M_k - \epsilon_k^{q+1}\left(1 + \exp(-a_k(\tau_k - b_k))\right)\right\}\right| < m_{th}$
  Calculate $E_{\text{sum}}^i = \sum_k E_k^{\text{Practical}}$, then let $i = i + 1, q = 0$.
end for
Find the maximum value among all E_{sum}^i, and the precoding and energy matrix.
Output: Use either of the methods to get the precoding vector $\mathbf{w}_k^{\text{opt}}$ and $\mathbf{V}^{\text{opt}}$.

We first simplify the objective function and then a new approximation will be formulated based on the *Bernstein-type Inequality* [77] [78]. By involving a simple transformation, we arrive at:

$$\mathbf{P}_9 : \max_{\mathbf{W}_k, \mathbf{V}, \rho} \sum_{k=1}^{K} \mu_k \left\{M_k - \epsilon_k\left(1 + \exp(-a_k(\tau_k - b_k))\right)\right\} \tag{5.37a}$$

$$\text{s.t.} \quad \Pr(E_k^{\text{In}} \geq \tau_k) \geq 1 - \varpi, \quad \forall \Delta\mathbf{h}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{H}_k), \forall k \in \mathcal{K}, \tag{5.37b}$$

$$(5.20b), (5.20e), (5.11e), (5.11f), (5.11g), (5.11h). \tag{5.37c}$$

Observe however that the transformation from (5.36a) to (5.37a) and (5.37b) is not exactly equivalent. The equivalent form should let $E_k^{\text{In}} \geq \tau_k$ in (5.37b). However, by setting ϖ to be a very small value, the transformation can be valid and it is also consistent with our gaussian CSI error model. By applying the *Bernstein-type Inequality I* [77], (5.37b)

becomes,

$$\mathrm{Tr}\big(\mathbf{H}_k^{1/2}\boldsymbol{\Sigma}\mathbf{H}_k^{1/2}\big) - \sqrt{-2\ln(\varpi)}\upsilon_{1k,s} + \ln(\varpi)\upsilon_{2k,s} + c_{k,s} \geq 0, \tag{5.38a}$$

$$c_{k,s} = \widehat{\mathbf{h}}_k^{\dagger}\boldsymbol{\Sigma}\widehat{\mathbf{h}}_k + \sigma_{k,S}^2 - \frac{\tau_k}{\rho}, \tag{5.38b}$$

$$\left\| \begin{bmatrix} \mathrm{vec}\big(\mathbf{H}_k^{1/2}\boldsymbol{\Sigma}\mathbf{H}_k^{1/2}\big) \\ \sqrt{2}\mathbf{H}_k^{1/2}\boldsymbol{\Sigma}\widehat{\mathbf{h}}_k \end{bmatrix} \right\| \leq \upsilon_{1k,s}, \tag{5.38c}$$

$$\upsilon_{2k,s}\mathbf{I} + \big(\mathbf{H}_k^{1/2}\boldsymbol{\Sigma}\mathbf{H}_k^{1/2}\big) \succeq \mathbf{0}, \upsilon_{2k,s} \geq 0, \ \forall k \in \mathcal{K}, \tag{5.38d}$$

where $\upsilon_{1k,s}$ and $\upsilon_{2k,s}$, $k \in \mathcal{K}$ are slack variables.

We also relax the problem by dropping the constraint that the rank of $\mathbf{W}_k$ must be 1, and the optimization problem becomes

$$\mathbf{P}_{10}: \max_{\substack{\mathbf{W}_k, \mathbf{V}, \rho, \{\upsilon_{1i,k}\} \\ \{\upsilon_{2i,k}\}, \{\upsilon_{1k,s}\} \\ \{\upsilon_{2k,s}\}, \{\upsilon_{1,n}\}, \{\upsilon_{2,n}\}}} \sum_{k=1}^{K} \mu_k\big\{M_k - \epsilon_k\big(1 + \exp(-a_k(\tau_k - b_k))\big)\big\},$$
$$\text{s.t. } (5.38), (5.23), (5.28), (5.11e), (5.11f), (5.11g). \tag{5.39a}$$

Still, the coupling variable in (5.38) can be tackled by fixing ρ. A similar one-dimensional search for ρ, together with a two-loop algorithm can solve $\mathbf{P}_{10}$, the detailed step will be omitted here for space considerations.

5.4.3 Complexity Analysis

For the CBS power minimization problem under the bounded CSI model, $\mathbf{P}_3$ has $\frac{K(K+1)}{2}$ linear matrix inequality (LMI) constraints of size $(M+1)$ in (13) due to the higher decoding complexity. Furthermore, we have N LMI constraints of size $(M+1)$ in (15) and K LMI constraints of size $(M+1)$ in (17). Additionally, in (12g), there are $(K+1)$ LMI constraints associated with size M, and a total of $\frac{K(K+1)}{2} + 2N + K + 2$ linear constraints.

Thus, according to [65] and the reference therein, the total complexity becomes

$$C_{\text{com}}^{B} = \ln(\tau^{-1})n\sqrt{\Psi_{\text{comp}}^{1}}\Bigg((\frac{K(K+1)}{2} + N + 2K + 1) [(M+1)^3 + n(M+1)^2] + (K+1)(M^3 + nM^2) + \frac{K(K+1)}{2} + 2N + K + 2 + n^2\Bigg), \tag{5.40}$$

where $n = \mathcal{O}\Big((K+1)M^2 + N + K + \frac{K(K+1)}{2}\Big)$, $\mathcal{O}$ is the big-O notation. Furthermore, we have $\Psi_{\text{comp}}^{1} = (\frac{K(K+1)}{2} + N + 2K + 1)M + K^2 + 4N + 3K + 4$, and τ is the accuracy of iteration.

Similarly, under Gaussian error model, there are $3K(\frac{K+3}{2}) + 3N + 2$ linear constraints, $\frac{K(K+1)}{2} + 2K + N + 1$ LMI of size M, and $\frac{K(K+1)}{2} + K + N$ second-order cone (SoC) constraints. Thus, the complexity becomes:

$$C_{\text{com}}^{G} = \ln(\tau^{-1})n\sqrt{\Psi_{\text{comp}}^{2}}\Bigg((\frac{K(K+1)}{2} + 2K + N + 1) [M^3 + nM^2] + 3K(\frac{K+3}{2}) + 3N + 2 + (\frac{K(K+1)}{2} + K + N)[(M^2 + M + 1)^2] + n^2\Bigg) \tag{5.41}$$

Where $\Psi_{\text{comp}}^{2} = 3K^2 + 10K + 6N + 3$.

For the maximum harvested energy problem, with bounded channel model, since the difference with that of power minimization problem is that a maximum of $T_{\max}$ number of iterations will be performed for one-dimensional search. Hence, the complexity is $T_{\max}C_{\text{com}}^{B}$. With Gaussian error model, the complexity is $T_{\max}^{'}C_{\text{com}}^{G}$, correspondingly, $T_{\max}^{'}$ is the number of unitary search.

5.5 Simulation Results

In this section, we present our simulation results for characterizing the performance of the proposed robust beamforming conceived with NOMA under both the bounded and the gaussian CSI estimation error models. Unless otherwise stated, the parameters are chosen

as in Table.5.1.

Parameters	Values
Number of SUs and PUs	$K = 3$, $N = 2$
Noise powers	$\sigma_{k,S}^2 = 0.1$, $\sigma_D^2 = 0.01$
Minimum required EH power	$P_{k,s} = 0.01$ Watt
Maximum tolerable interference of PUs	$P_{n,p} = -18$ dBm
Estimated channel gains	$\hat{\mathbf{h}}_k \sim \mathcal{CN}(\mathbf{0}, 0.8\mathbf{I})$ $\hat{\mathbf{g}}_n \sim \mathcal{CN}(\mathbf{0}, 0.1\mathbf{I})$
Outage probability threshold	$\xi_k = \xi_{k,s} = \xi_{n,p} = 0.05$
Gaussian CSI estimation	$\varpi_k^2 = 0.001$, $\varpi_n^2 = 0.0001$ [65]
Non-linear EH model	$M_k = 24$ mW, $a_k = 150$ $b_k = 0.014$ [80]

Table 5.1: Simulation parameters for chapter 5

To achieve a fair comparison between the two channel estimation error models. If the covariance matrices of the channel estimation error vector $\Delta\mathbf{h}_k$ and $\Delta\mathbf{g}_n$ under the gaussian model are $\varpi_k^2\mathbf{I}$ and $\varpi_n^2\mathbf{I}$, respectively, then the bounded CSI radius under the worst-case scenario of φ_k and ψ_n should be [75]

$$\varphi_k = \sqrt{\frac{\varpi_k^2 F_{2M}^{-1}(1-\xi_k)}{2}},\ \psi_n = \sqrt{\frac{\varpi_n^2 F_{2M}^{-1}(1-\xi_{n,p})}{2}}, \tag{5.42}$$

where $F_{2M}^{-1}(\cdot)$ represents the complimentary cumulative distribution function (CCDF) of the Chi-square distribution with $2M$ degrees of freedom.

5.5.1 Power Minimization Problem

Fig. 5.2 shows the empirical CDFs of the minimum transmit power of the CBS for both the imperfect CSI estimation error models. The maximum power P_B is set to 2 Watts. For comparison, we also include the case of OMA, since it represents the traditional access technology. Observe that in order to reduce the inter-user interference, each OMA user relies exclusively on a single time slot. Thus, a total of K time slots are required instead of a single one in our scheme. To make a fair comparison, each SU's achievable data rate

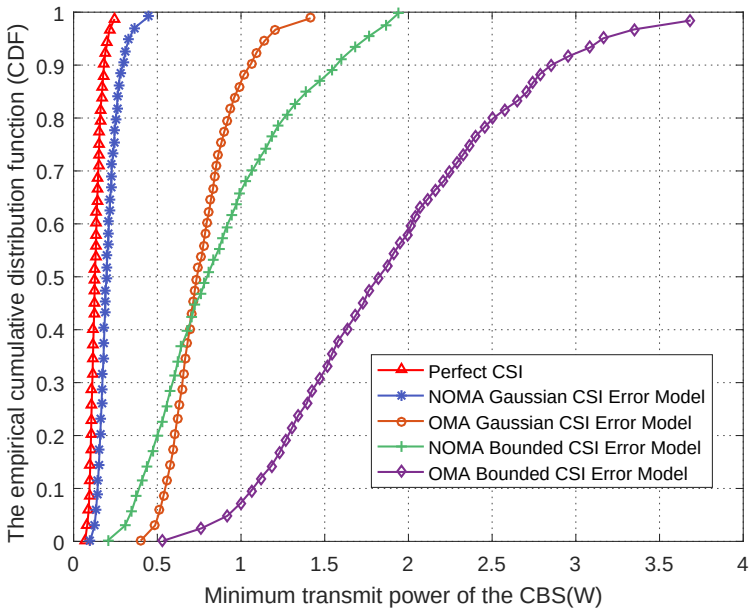

Fig. 5.2: The empirical CDF of the minimum transmit power of the CBS under different channel conditions. CBS antenna number $M = 10$, $P_B = 2$ Watts, $R_{\min} = 1$ bit/s/Hz.

should be averaged over all K time slots, which becomes $R_k^{\mathrm{OMA}} = \frac{1}{K}\log_2(1 + \mathrm{SINR}_k^{\mathrm{OMA}})$. Reduced interference is achieved at the cost of a lower spectral and energy efficiency. We also observe that under both channel error models, the performance of NOMA is better than that of OMA. This is because for OMA, the lower spectral efficiency makes the SU data rate requirement harder to be satisfied. Hence the CBS has to apply a higher transmission power to compensate for that, which leads to a much higher energy consumption. Fig. 5.2 is generated from 1,000 independent realizations of different channel conditions. As expected, the performance under perfect CSI is the best, since no additional power is used to compensate for the channel uncertainties. Furthermore, in both the OMA and NOMA schemes, the performance under the gaussian CSI channel estimation is better than that under the bounded CSI channel estimations, as bounded CSI represents the worst-case scenario. Observe that the minimum power in the OMA bounded CSI is over 2 Watts since we only limit the power of each time slot to 2 Watts and it is very likely that the total power over K slots will beyond that limit.

Fig. 5.3 shows the minimum transmit power of the CBS as a function of the minimum

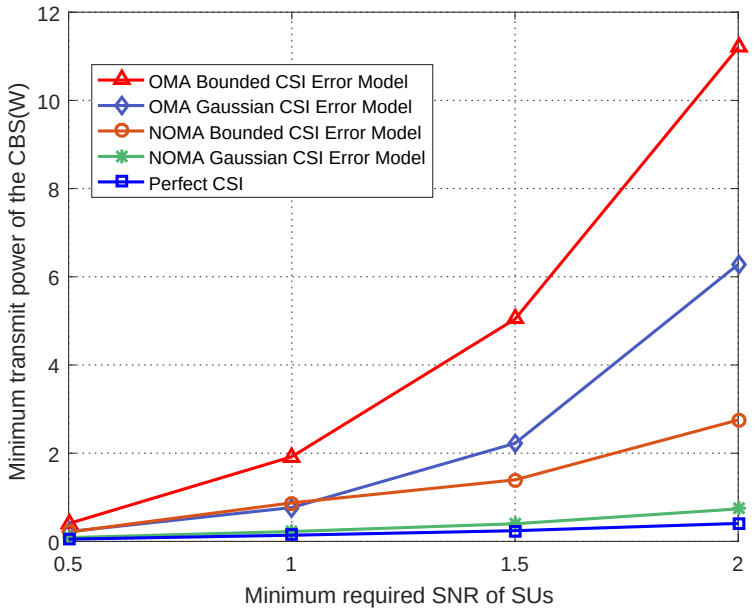

Fig. 5.3: The minimum transmit power of the CBS vs. the required SNR of SUs for $M = 10$, $P_B = 8$ Watts.

required SNR of SUs, $\gamma_{k,\min}$. As the SNR increases, the power increases under all CSI cases. Also, perfect CSI requires the least power, followed by NOMA relying on the gaussian CSI error model, NOMA in the bounded CSI model, OMA gaussian CSI model, and OMA bounded CSI model. Besides, compared to OMA, the CBS power in NOMA grows more slowly. In the parameter setting, $\gamma_{k,\min}$ plays a more important role in the constraints. For $\gamma_{k,\min} = 2$ in the NOMA case, the equivalent SNR for OMA will be 26. Thus, the gap between OMA and NOMA further increases with the required SNR.

The impact of the CBS antenna number is illustrated in Fig. 5.4(a), where the performance with different CBS antenna numbers and channel uncertainties are plotted. Specifically, Fig. 5.4(a) illustrates how the number of antennas affects the overall performance. The power required increases, when the SNR of SUs grows, regardless of how many antennas are mounted at the CBS. It is also observed that the minimum power required decreases when the number of antennas increases, since a larger number of antennas results in a higher degree of freedom (DoF). Besides, we also notice that the performance under the gaussian error model is better than that under the bounded channel error case. In Fig.

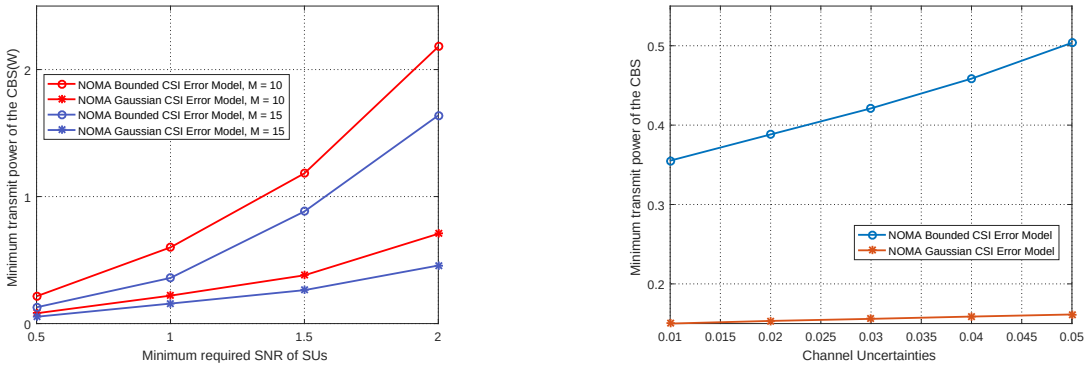

Fig. 5.4: (a) Impact of the number of CBS antennas on the minimum transmitted power required in two imperfect CSI scenarios. (b) Impact of channel uncertainties ψ_n and φ_k on the overall minimum transmit power of the CBS, $M = 15$, $R_{\min} = 1$ bit/s/Hz, $P_B = 8$ Watts.

5.4(b), the impact of channel uncertainties is illustrated. We set $\psi_n^2 = \varphi_k^2 = [0.01 : 0.05]$, the corresponding covariance matrices in gaussian CSI estimation error scenario also change according to (5.42). Clearly, channel estimation error affects the bounded CSI scenario the most, since under worst-case CSI, the channel estimation error channel becomes worse, thus it needs more power to meet the data rate constraints. Nevertheless, the channel estimation error does not have much impact on the gaussian channel estimation error scenario.

5.5.2 Energy Harvesting Maximization Problem

In this subsection, we present results for the maximum EH as our objective function. The CBS power is $P_B = 2$ Watts. Fig. 5.5 characterizes the average maximum EH power vs. the interference tolerated by the to PUs. One can observe that the energy harvested monotonically increases, when the maximum interference tolerated by the PUs grows, where a higher $P_{n,p}$ allows for a larger transmission power, leading to the increase of the harvested energy. Additionally, we can see that under the gaussian channel estimation error, the performance is better than that under the bounded channel estimation error case. When the channel conditions are better, less power is required for satisfying the data rate requirements. Hence more power can be reserved for EH. This also explains that when the required SNR is low, a high EH power can be achieved.

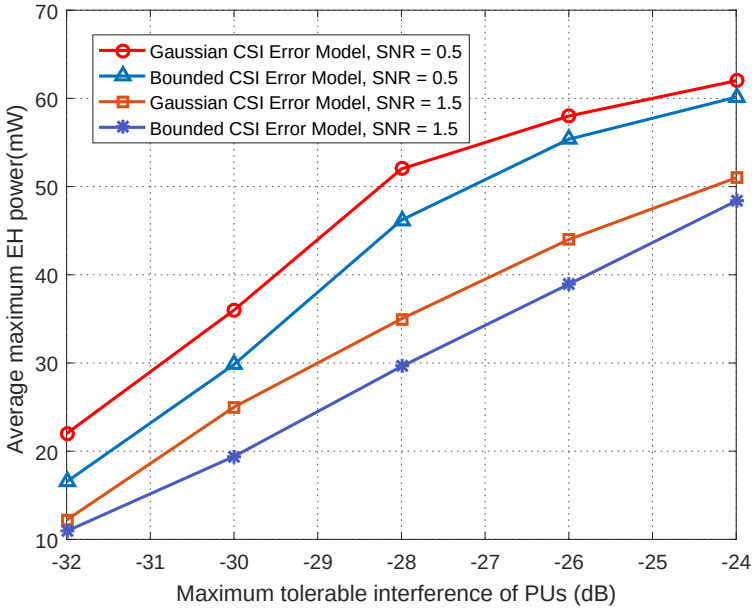

Fig. 5.5: Average maximum EH power under different interferences tolerated by the PUs, $M = 10$.

The impact of minimum SNRs required by the SUs is illustrated in Fig. 5.6. The number of CBS antennas is $M = 10$ and the interference threshold $P_{n,p}$ is set to -24 dBm. We also list the results for the OMA cases. As expected, the average maximum EH power decreases, when the required SNR increases. Similar observations show that under perfect CSI, the performance is the best, while the OMA bounded CSI estimation scenario is the worst. Moreover, we can see that the maximum EH power decreases significantly when the SNR grows. This is because more power has to be used for information detection, which leaves less power for energy harvesting.

Fig. 5.7 shows the average total EH power vs. the number of SUs. It can be observed that the total EH power grows, when the number of SUs increases, since more nodes participate in the harvesting process. Additionally, we can see that when the number of antennas is higher, more EH power can be achieved. This is because more antennas give a higher system DoF, therefore less power is sufficient for information detection.

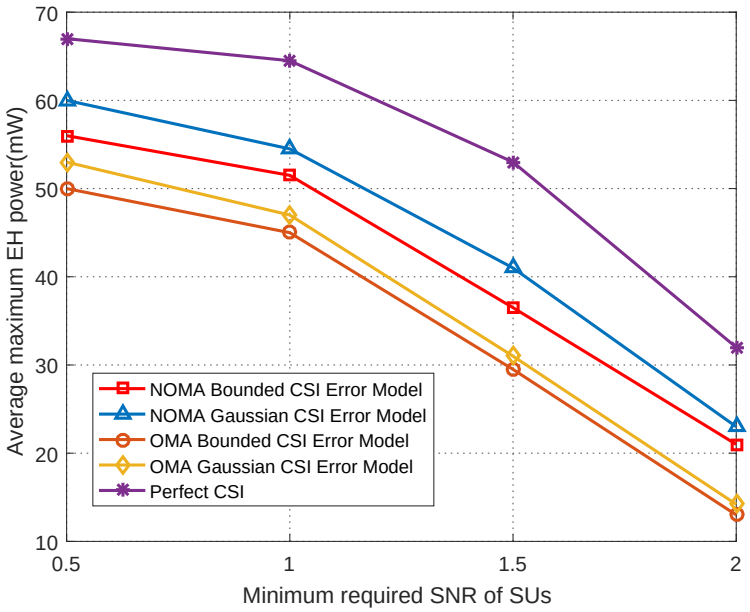

Fig. 5.6: Average maximum EH power vs. the minimum SNR required by the SUs, $M = 10$.

5.6 Chapter Conclusions

In this chapter, we considered MISO-NOMA CR-aided SWIPT under both the bounded and the Gaussian CSI estimation error model. To make the energy harvesting investigations more realistic, a non-linear EH model was applied. Robust beamforming and power splitting control were jointly designed for achieving the minimum transmission power and maximum EH. We transformed the non-convex minimum transmission power optimization problems into a convex form while applying a one-dimensional search algorithm to solve the maximum EH problem. Our simulation results showed that the performance achieved by using NOMA is better than that obtained by using the traditional OMA. Furthermore, a performance gain can be obtained under the gaussian CSI estimation error model over the bounded CSI error model. As for future research directions, the system model can be generalized to account for more use cases, for example, considering the physical layer security and the interference arising from multipoe cells. Additionally, for the Gaussian CSI error model, the rank of the solution is not fully characterized in this work.

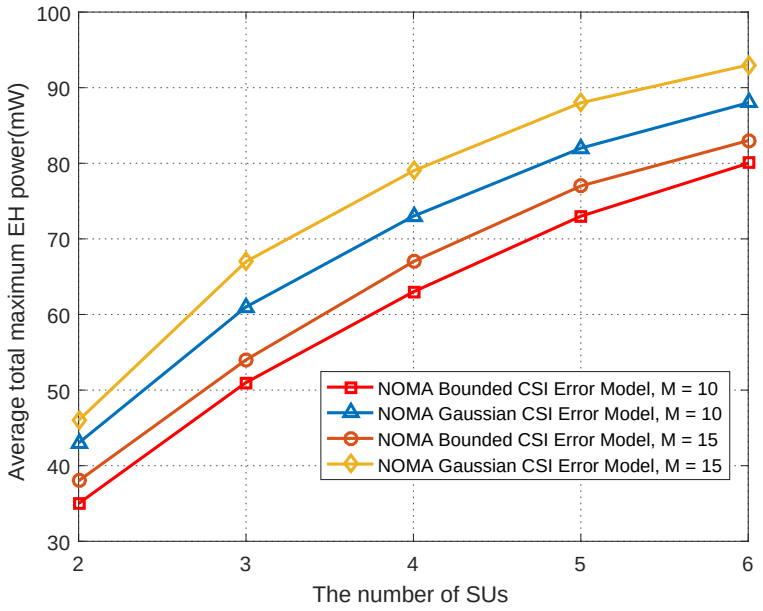

Fig. 5.7: Average total EH power vs. the number of SUs for $P_{n,p} = -24$ dBm, $r_{\min} = 1$ bit/s/Hz.

CHAPTER 6

Joint Offloading and Computation Energy Efficiency Maximization in a Mobile Edge Computing System

6.1 Introduction

In the previous chapters, we mainly addressed the spectral and large number of device connection challenges in the future 5G systems. However, the ever increasing demand for various applications such as gaming, autonomous driving, and AR/VR, have been recognized as one of the driving forces for the prosperity of the smart devices [12]. Due to the limitations on size, battery, and cost, these small size smart devices can experience performance bottleneck when computation-intensive tasks need to be executed. One option is to deploy centralized services such as cloud centers to help the data processing. However, cloud servers can be located far away, which can inevitably cause longer end-to-end transmission delay [12].

In contrast to the centralized infrastructure, recent network paradigms such as MEC tend to allocate resources to devices in close-proximity for joint processing. For example, the work in [81] used unmanned aerial vehicles (UAVs) to help D2D wireless networks [82]. This paradigm shift can effectively reduce the long backhaul latency and energy consumption, as well as support a more flexible infrastructure in a cost-effective way. Furthermore, MEC together with virtual machine (VM) migration can effectively increase the scalability [83] while reduce service delay [84]. Due to these advantages, MEC has attracted extensive research attentions in various vertical segments.

One important feature of MEC is performing computation offloading, which leverages the powerful MEC servers in proximity and sends the computation-intensive tasks to MEC servers for processing. It can help overcome the physical limitations of local small devices. Current research involves two categories of offloading: binary [85] and partial [86]- [89].

Binary offloading executes the task as a whole, either locally or in the MEC server, while partial offloading assumes the task can be partitioned into two parts, one for local processing and one for offloading. Even though the former is easier in implementation, for a very large dataset, partial offloading can help reduce the latency and energy consumption on the local devices more effectively.

Previous works either target on minimizing the total energy consumption or maximizing total computed bits. Energy-efficient communication has received tremendous industrial and academic attention in various systems such as multi-hop and heterogeneous networks [27]. By applying energy efficiency as the performance metric, QoS can be obtained, together with a reduction on energy consumption [90]. Energy efficiency defined in traditional communication systems in bits transmitted per Joule is an important metric to evaluate the overall system energy consumed. However, in the new communications systems, there exist a large number of computation constrained and power limited devices (such as IoT devices) that will need to support delay-critical yet computation-intensive tasks. Offloading through communications to MEC servers in order to compute the tasks timely becomes critically important to meet the short delay budget requirement while communications throughput requirements may become secondary. To capture the efficiency of energy used for both computing and communication in such a scenario, we propose the metric computation efficiency, which is defined as the number of total computed bits divided by the energy consumed. We argue that this metric is more appropriate since it can measure how efficient the system is, in terms of computed bits per Joule, for a system involving massive computation needs.

Our work expands [87] and [90] in two major aspects. Firstly, we consider maximizing the computation efficiency instead of purely maximizing computed data bits or minimizing energy consumption compared with [87]. Secondly, we combine local computing and data offloading in a hybrid approach instead of offloading only [90]. The contributions of this paper are briefly summarized as follows.

1. We propose a new performance metric in MEC networks: computation efficiency, which is defined as the number of computed bits divided by the corresponding ener-

gy consumption. Computation efficiency can drive towards efficient on-board power utilization while achieving satisfactory QoS.

2. The fundamental trade-off between local computing and data offloading is analyzed. Results show that with practical parameter settings, when data size is small, more data will be processed locally. But when the data grows, offloading will play a more important role in improving the computation efficiency.

6.2 System Model

In this chapter, we consider a downlink MEC network which consists of one MEC server and K randomly located UEs. The server has a single antenna and so does each UE. Assume the channel between the server and the UE is a block-fading-based model, i.e., the channel remains constant during a time slot with length T but varies from time to time. The channel state information is assumed to be available at the server. At the beginning of a particular time T, each UE has a computation-intensive task to compute. Due to the computation resource limit or power limit or both, these tasks are offloaded to the nearby MEC server for a more powerful processing if needed. In this article, we assume the task-input bits are bit-wise independent and can be arbitrarily divided into different groups and executed by different entities in MEC system, e.g., parallel execution at the mobile and MEC server [87]. Partial offloading is used here. Thus the system can support data offloading and local computing simultaneously.

6.2.1 Data Offloading

Denote the set of UEs as $\mathcal{K} = \{1, 2, \ldots, K\}$. A UE can offload part of the computation bits to the server. To reduce the interference between different UEs, UEs doing offloading are allocated a portion of T and transmit sequentially, such as in the TDMA mode. Specifically, let g_k, p_k, and t_k respectively represent the channel between the server and UE k, the transmission power, and time duration allocated to UE k. The total number of offloaded bits is $r_k = B \log_2 \left(1 + \frac{p_k g_k}{\sigma^2}\right) t_k, \ \forall k \in \mathcal{K}$, where σ^2 is the noise power and B is the system

bandwidth.

Under this mode, the corresponding energy consumption for UE k is $e_k = p_k t_k + p_r t_k$, where $p_k t_k$ denotes the over-the-air information transmission energy consumption, and p_r is the constant circuit power for transmit signal processing, which is the same for all UEs.

6.2.2 Local Computing

In addition to offloading, part of the bits can be computed locally by UEs. Let C_k be the number of computation cycles needed to process one bit of data for UE k. Clearly, each UE can compute the data throughout the entire block T. Furthermore, f_k denotes the processor's computing speed in the unit of cycles per second, and similar to [85], this speed holds constant. Therefore, the total number of bits locally computed is $r_k^{\text{local}} = \frac{T f_k}{C_k}$. The energy consumption of local computing is modeled as a function of the processor speed f_k. Specifically, $E_k^{\text{local}} = \epsilon_k f_k^3 T$, where ϵ_k is the computation energy efficiency coefficient of the processor's chip [85] [91].

6.3 Problem Formulation

In this section, we form an optimization problem that maximizes the total computation energy efficiency among all UEs. Mathematically, the problem is expressed as follows.

$$\mathbf{P}_1 \max_{\{t_k\},\{f_k\},\{p_k\}} \sum_k w_k \frac{B \log_2(1 + \frac{p_k g_k}{\sigma^2}) t_k + \frac{T f_k}{C_k}}{\epsilon_k f_k^3 T + p_k t_k + p_r t_k} \tag{6.1a}$$

$$\text{s.t. } C1 \quad : \quad \sum_k t_k \leq T, \tag{6.1b}$$

$$C2 \quad : \quad B \log_2(1 + \frac{p_k g_k}{\sigma^2}) t_k + \frac{T f_k}{C_k} \geq L_k, \forall k, \tag{6.1c}$$

$$C3 \quad : \quad \epsilon_k f_k^3 T + p_k t_k + p_r t_k \leq E_k^{th}, \forall k, \tag{6.1d}$$

$$C4 \quad : \quad 0 \leq f_k \leq f_k^{\max}, \forall k, \tag{6.1e}$$

$$C5 \quad : \quad t_k \geq 0, \forall k, \tag{6.1f}$$

where w_k is the weighting factor that can be used to prioritize different QoS requirements of UEs. $\mathbf{P}_1$ is a resource allocation problem that optimizes the offloading transmission time

t_k and power p_k, as well as local computing chip frequency f_k. C_1 states that all the tasks should be completed before the end of the block. Notice that here we omit the processing and transmission time at the server by following [85] [87]. L_k in C_2 denotes the minimum data bits for computing for UE k. E_k^{th} in C_3 is the total energy available in UE k. C_4 defines the maximum CPU frequency of each UE.

The above problem is non-convex since the objective function involves sum-of-ratio maximization. Also, the coupling of some variables makes the optimization problem even more complicated. To address the coupling problem, let $P_k = p_k t_k$. Besides, for notational brevity, denote $R_k(P_k, t_k, f_k) = B\log_2(1+\frac{P_k g_k}{t_k \sigma^2})t_k + \frac{T f_k}{C_k}$, and $E_k(P_k, t_k, f_k) = \epsilon_k f_k^3 T + P_k + p_r t_k$. We first employ simple transformations and the original problem becomes:

$$\mathbf{P}_2: \quad \max_{\{t_k\},\{f_k\},\{P_k\},\{\beta_k\}} \sum_k w_k \beta_k \tag{6.2a}$$

$$\text{s.t. } C1: \quad R_k(P_k, t_k, f_k) \geq \beta_k E_k(P_k, t_k, f_k), \tag{6.2b}$$

$$C2: \quad \sum_k t_k \leq T, \tag{6.2c}$$

$$C3: \quad t_k \geq 0, \forall k, \tag{6.2d}$$

$$C4: \quad 0 \leq f_k \leq f_k^{\max}, \forall k, \tag{6.2e}$$

$$C5: \quad R_k(P_k, t_k, f_k) \geq L_k, \tag{6.2f}$$

$$C6: \quad E_k(P_k, t_k, f_k) \leq E_k^{th}, \forall k. \tag{6.2g}$$

Lemma 1: For $\forall k$, if $(\{t_k^*\}, \{f_k^*\}, \{P_k^*\}, \{\beta_k^*\})$ is the optimal solution of $\mathbf{P}_2$, there must exist $\{\lambda_k^*\}$ such that $(\{t_k^*\}, \{f_k^*\}, \{P_k^*\})$ satisfies the Karush-Kuhn-Tucker condition of the following problem for $\lambda_k = \lambda_k^*$ and $\beta_k = \beta_k^*$.

$$\mathbf{P}_3: \quad \max_{\{t_k\},\{f_k\},\{P_k\}} \sum_k \lambda_k (w_k R_k - \beta_k E_k) \tag{6.3a}$$

$$\text{s.t.} \quad (6.2c) - (6.2g). \tag{6.3b}$$

Furthermore, $(\{t_k^*\},\{f_k^*\},\{P_k^*\})$ satisfies the following equations for $\lambda_k=\lambda_k^\star$ and $\beta_k=\beta_k^\star$:

$$\lambda_k=\frac{1}{E_k(P_k,t_k,f_k)},\ \beta_k=\frac{w_kR_k(P_k,t_k,f_k)}{E_k(P_k,t_k,f_k)},\forall k. \tag{6.4}$$

Lemma 1 can be proved by taking the derivative of the Lagrange function of $\mathbf{P}_2$. λ_k is the non-negative multiplier of (6.2b). A detailed proof can be obtained in [79]. *Lemma 1* implies that the optimal solution of $\mathbf{P}_2$ can be obtained by solving the equations of (6.4) among the solutions of $\mathbf{P}_3$.

The Lagrange function of $\mathbf{P}_3$ is

$$\begin{aligned}&\mathcal{L}(t_k,P_k,f_k,\alpha_k,\mu_k,n_k,m) \\ =&\sum_k\lambda_k(w_kR_k-\beta_kE_k)-\sum_k\alpha_k(E_k-E_k^{th}) \\ -&\sum_k\mu_k(L_k-R_k)-\sum_k n_k(f_k-f_k^{\max})-m(\sum_k t_k-T),\end{aligned} \tag{6.5}$$

where $\alpha_k,\mu_k,\theta_k,n_k$, and m are non-negative Lagrange multipliers for the respective constraints. It can be readily proved that $\mathbf{P}_3$ is convex for given λ_k and $\beta_k,\forall k$, and satisfies Slater's condition. Thus, strong duality holds between the primal and dual problems, which means solving $\mathbf{P}_3$ is equivalent to solving the dual problem. Notice that the dual function is $\psi(\alpha_k,\mu_k,n_k,m)=\max\limits_{\{t_k\},\{f_k\},\{P_k\}}\mathcal{L}(t_k,P_k,f_k,\alpha_k,\mu_k,n_k,m)$. The dual problem becomes

$$\mathbf{P}_4:\min_{\alpha_k,\mu_k,n_k,m}\psi(\alpha_k,\mu_k,\theta_k,n_k,m). \tag{6.6}$$

In the following, we first obtain the optimal solutions for the given auxiliary variables (λ_k,β_k) and Lagrange multipliers (α_k,μ_k,n_k,m). Then the Lagrange multipliers are updated via gradient descent method. Lastly, the auxiliary variables are updated as well.

6.3.1 Update p_k, t_k, and f_k

Equation (6.5) can be re-organized as

$$\begin{aligned} & \mathcal{L}(t_k, P_k, f_k, \alpha_k, \mu_k, n_k, m) \\ = & \sum_k \Big((\lambda_k w_k + \mu_k) R_k - (\alpha_k + \lambda_k \beta_k) E_k - n_k f_k - m t_k \\ + & \alpha_k E_k^{th} - \mu_k L_k + n_k f_k^{\max} \Big) + mT. \end{aligned} \tag{6.7}$$

To maximize the dual function, $\psi(\alpha_k, \mu_k, \theta_k, n_k, m)$ can be decomposed into K sub-problems. Specifically, the k-th problem is

$$\begin{aligned} \psi_k &= \max_{\{t_k\},\{f_k\},\{P_k\}} \mathcal{L}_k(t_k, P_k, f_k, \alpha_k, \mu_k, n_k, m) \\ &= \max_{\{t_k\},\{f_k\},\{P_k\}} (\lambda_k w_k + \mu_k) R_k - (\alpha_k + \lambda_k \beta_k) E_k - n_k f_k - m t_k + \Psi, \end{aligned} \tag{6.8}$$

where Ψ denotes the constant value that is irrelevant to the optimizing variables.

Proposition 1. *The optimal transmit power and duration for the k-th UE should be* $p_k^* = \left[\frac{(\lambda_k w_k + \mu_k) B}{(\lambda_k \beta_k + \alpha_k) \ln 2} - \frac{\sigma^2}{g_k}\right]^+$ *and* $f_k^* = \sqrt{\left[\frac{\left(\frac{(\lambda_k w_k + \mu_k)}{c_k} - n_k\right) \frac{1}{\epsilon_k}}{3(\lambda_k \beta_k + \alpha_k)}\right]^+}$ *respectively, where* $[x]^+ = \max(x, 0)$.

Proof. Taking the derivative of the Lagrange function ψ_k w.r.t. P_k yields

$$\frac{\partial \psi_k}{\partial P_k} = \frac{(\lambda_k w_k + \mu_k) B t_k g_k}{(t_k \sigma^2 + P_k g_k) \ln 2} - \lambda_k \beta_k - \alpha_k. \tag{6.9}$$

Let $\frac{\partial \psi_k}{\partial P_k} = 0$, the optimal P_k^* can be obtained. Notice that the optimal p_k^* is equal to $\frac{P_k^*}{t_k}$. Similarly, let $\frac{\partial \psi_k}{\partial f_k} = 0$, we can get the optimal expression for f_k. □

- *Remark:* In order to maximize EE, user k with a higher channel gain g_k should transmit with a higher power p_k. This can be seen from the optimal expression of p_k^*. Notice that the similar conclusion is also drawn in [90].

For t_k, the partial derivative expression of ψ_k w.r.t. t_k becomes

$$\frac{\partial \psi_k}{\partial t_k} = (\lambda_k w_k + \mu_k) B \log_2(1 + \frac{p_k g_k}{\sigma^2}) - (\alpha_k + \lambda_k \beta_k)(p_k + p_r) - m. \tag{6.10}$$

Clearly, the optimization problem is a linear function of t_k. Therefore, the following problem can be solved efficiently by interior point methods.

$$\mathbf{P}_5: \quad \max_{\{t_k\}} \sum_k \lambda_k (w_k R_k - \beta_k E_k) \tag{6.11a}$$

$$\text{s.t.} \quad (6.2c), (6.2d), (6.2f), (6.2g). \tag{6.11b}$$

6.3.2 Update Lagrange Multipliers

Now, we proceed to update the Lagrange multipliers α_k, μ_k, n_k, and m. From the problem definition, with known P_k, t_k, and f_k, the dual problem is always convex. Specifically, $\min_{\substack{\alpha_k,\mu_k,\\ \theta_k,n_k,m}} \psi(\alpha_k, \mu_k, \theta_k, n_k, m)$ is an affine function w.r.t. dual variables. Thus, we can apply the simple gradient method for the variable update. Specifically, we choose initial $\alpha_k(0), \mu_k(0), n_k(0)$, and $m(0)$ as the center of the ellipsoid which contains the optimal Lagrange variables. Then, we reduce the volume of the ellipsoid using gradient descent method as the following.

$$\alpha_k(i+1) = \alpha_k(i) + \Delta\alpha_k(E_k^* - E_k^{th}), \tag{6.12a}$$

$$\mu_k(i+1) = \mu_k(i) + \Delta\mu_k(L_k - R_k^*), \tag{6.12b}$$

$$n_k(i+1) = n_k(i) + \Delta n_k(f_k^* - f_k^{\max}), \tag{6.12c}$$

$$m(i+1) = m(i) + \Delta m(\sum_k t_k^* - T), \tag{6.12d}$$

where $\Delta\alpha_k, \Delta\mu_k, \Delta n_k$, and Δm are the respective step size, i is the iteration index. Notice that all the Lagrange variables must be non-negative. If a negative value is obtained, the Lagrange variable will be set to 0 instead.

6.3.3 Update Auxiliary Variables

Lastly, the auxiliary variables λ_k and β_k are updated in the following way.

Notice that in *Lemma 1*, the optimal solution P_k^*, t_k^*, and f_k^* should also satisfy the following system conditions:

$$\beta_k E_k(P_k^*, t_k^*, f_k^*) - w_k R_k(P_k^*, t_k^*, f_k^*) = 0, \tag{6.13}$$

$$\lambda_k E_k(P_k^*, t_k^*, f_k^*) - 1 = 0. \tag{6.14}$$

Similarly, according to [**?**], we define functions for notational brevity. Specifically, let $T_j(\beta_j) = \beta_j E_k - w_k R_k$ and $T_{j+K}(\lambda_j) = \lambda_j E_k - 1$, $j \in \{1, 2, \ldots, K\}$. The optimal solution for λ_k and β_k can be obtained by solving $\mathbf{T}(\lambda_k, \beta_k) = [T_1, T_2, \ldots, T_{2K}] = \mathbf{0}$. We can apply iterative method to update the auxiliary variables. Specifically,

$$\lambda_k(i+1) = (1 - \theta(i))\lambda_k(i) + \frac{\theta(i)}{E_k(P_k^*, t_k^*, f_k^*)}, \tag{6.15}$$

$$\beta_k(i+1) = (1 - \theta(i))\beta_k(i) + \theta(i)\frac{w_k R_k(P_k^*, t_k^*, f_k^*)}{E_k(P_k^*, t_k^*, f_k^*)}, \tag{6.16}$$

where $\theta(i)$ is the largest θ that satisfies $||\mathbf{T}(\lambda_k(i) + \theta^l \mathbf{q}^i_{K+1:2K}, \beta_k(i) + \theta^l \mathbf{q}^i_{1:K})|| \leq (1 - z\theta^l)||T(\lambda_k(i), \beta_k(i)||$, $\mathbf{q}$ is the Jacobian matrix of $\mathbf{T}$, $l \in \{1, 2, \ldots\}$, $\theta_l \in (0, 1)$, and $z \in (0, 1)$. Note that when $\theta(i) = 1$, it becomes the standard Newton method. To summarize, we list the detailed algorithm in **Algorithm 1**.

Algorithm 3 Computation Efficiency Maximization Algorithm

Initialization: the algorithm accuracy indicator t_1 and t_2, set $i = 0$, $\lambda_k(i)$ and $\beta_k(i)$
while $||\mathbf{T}(\lambda_k, \beta_k)|| > t_1$ **do**
  Initialization: $\alpha_k(j), \mu_k(j), n_k(j)$ and $m(j)$, and let $j = 0$
  while $|\alpha_k(j+1) - \alpha_k(j)| > t_2$ **do**
    Calculate p_k^* and f_k^* based on *Proposition 1*.
    Solve for problem $\mathbf{P}_5$, obtain the timing variable t_k.
    Update Lagrange variables based on gradient descent method in (6.12).
    Let $j = j + 1$.
  end while
  Let $i = i+1$, update auxiliary variables $\lambda_k(i+1)$ and $\beta_k(i+1)$ from (6.15) and (6.16).
end while
Output the optimal computation efficiency.

Notice that in the inner loop, the stop criterion can also be the convergence of other Lagrange multipliers or the condition that their combined value is less than a threshold.

6.3.4 Complexity Analysis

Since the algorithm involves the iteration process for three variables, we analyse the complexity in a sequential way. Firstly, p_k and f_k have a linear complexity with the user number K. The updating of Lagrange variables is of $\mathcal{O}(K^2)$ complexity since the total number of variables are $3K + 1$. Here $\mathcal{O}(x)$ means the upper bound for the complexity grows with order x. Finally, auxiliary variables λ_k and β_k have a complexity independent of K. Thus, our proposed algorithm has a total complexity in $\mathcal{O}(K^3)$.

6.4 Performance Evaluation

In this section, we present our simulation results of the joint offloading and computation scheme. The parameters are set as follows. The system bandwidth is $B = 200$ kHz, block length $T = 1s$, total number of UEs $K = 2$, $C_k = 10^3$ cycles needed for one bit raw data processing, the chip computing efficiency $\epsilon_k = 10^{-24}$, and static circuit power $p_r = 50$ mW. The channels between the MEC server and each UE are modelled as the joint effect of large-scale and small-scale fading, with $g_k/\sigma^2 = G_k h_k$, $G_1 = 7$, and $G_2 = 3$. h_k is the unitary Gaussian random variable. Lastly, the maximum computation capacity of each UE is set equally as $f_k^{\max} = 10^9$ Hz. $E_1^{th} = E_2^{th} = 2$ Joule. All the results are averaged over different random channel realizations.

In Fig. 6.1, we present the comparison results among three schemes, namely, the proposed scheme in this paper, offloading only scheme, and local computing only scheme. We set $L_1 = L_2$ and $w_1 = w_2 = 1$, which means the minimum required data bits for all UEs are the same. In Fig. 6.1, the computation efficiency of all the schemes decreases with the increase of the minimum required data bits. This suggests that the energy required to compute grows faster than the growth of the data bits. It is evident that our proposed algorithm outperforms other schemes. Additionally, we notice that when the data size is small, the proposed scheme's performance is closer to that of local computing only; when the

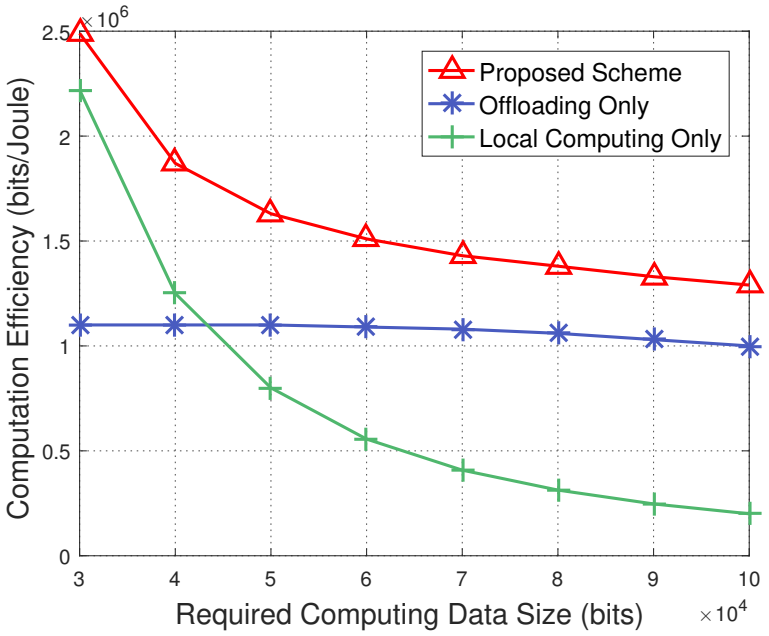

Fig. 6.1: Performance comparison of different schemes

data size grows, the performance will approach to that of offloading only. This phenomenon can be explained by the following. Firstly, in the real world applications, processor clock speed in a mobile device can reach MHz level. Thus, when the data size is relatively small, the preferred choice is to compute locally. Furthermore, based on the channel gain between UEs and the BS, and also the available bandwidth, data offloading may not be the ideal choice since it may take a longer time and a higher energy for small data offloading than for small data computing locally. On the other hand, when the data size is large, offloading to more computation powerful MEC server can become a much better choice. Moreover, the energy decrease in local computing is more dramatic than the energy used in offloading when data size shrinks. According to the equation for local computing only, the computation efficiency is $\frac{r_k^{\text{local}}}{E_k^{\text{local}}} = \frac{1}{C_k \epsilon_k f_k^2} \propto \frac{1}{r_k^2}$, which indicates that its efficiency is inversely proportional to the square of the data size, while the offloading has a much slower deceasing rate thanks to the *log* function.

Compared with partial offloading, another MEC offloading scheme is the binary offloading, where each UE either completely offloads all the data to the MEC server or computes

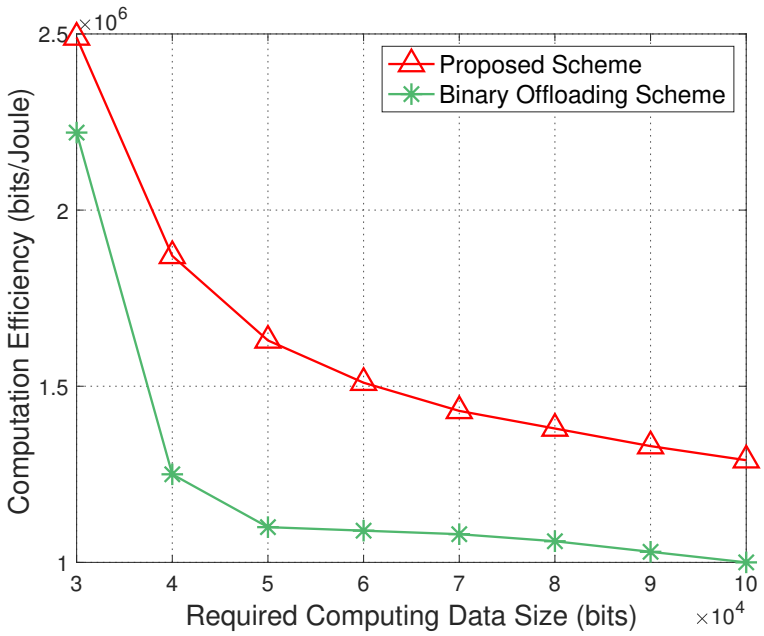

Fig. 6.2: Performance comparison of our proposed scheme and the binary offloading

all the data locally. To compare its performance with our proposed scheme, we show the result in Fig. 6.2. It can be seen that our proposed joint scheme outperforms the binary offloading in terms of computation efficiency, which indicates the superiority of the proposed algorithms.

Fig. 6.3 illustrates the trade off between two strategies: data offloading and local computing in our proposed scheme. The vertical axis represents the number of data bits (in percentage) calculated by either scheme with respect to the whole task. It can be readily shown that for both UEs, the local computing amount (in percentage) will decrease with the increase of the preset data amount. By contrast, data offloading plays a more and more important when the data become large. This can further prove our point in Fig. 6.1, where the proposed scheme adaptively adjusts the amount of data that go through local computing or offloading. Additionally, for UE 1, the trade off point happens around $L_1 = 4 \times 10^4$ and for UE 2 around $L_2 = 6 \times 10^4$. Since UE 1 has a better channel gain than UE 2, the influence from data offloading is more prominent, thus the trade off point is in an earlier position while for UE 2, local computing continues to have a more influential role until the

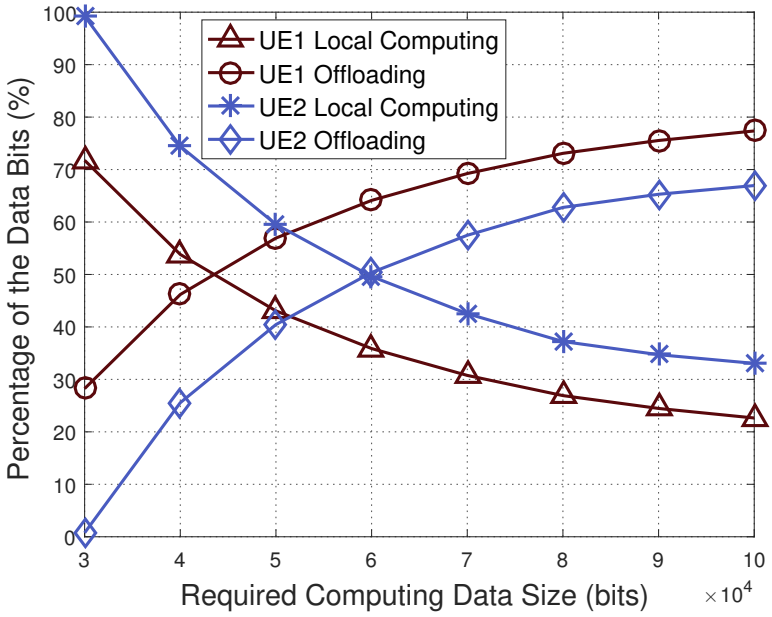

Fig. 6.3: Trade off between offloading and local computing

trade off point $L_2 = 6 \times 10^4$.

6.5 Chapter Conclusions

In this chapter, we present a new evaluation metric in MEC systems, i.e. the computation efficiency. An optimization problem is formulated which aims at maximizing the total computation efficiency with weight factors. The problem is recognized as the sum-of-ratio problem and an iterative algorithm is applied in the outer loop. For the inner loop, the problem can be converted to standard convex optimization and to gain a better insight, we propose to solve it via gradient descent method. Simulation results reveal the fundamental trade-off of two combined schemes: local computing and offloading.

CHAPTER 7

Wearable Communications in 5G: Challenges and Enabling Technologies

7.1 Introduction

The upcoming 5G aims to support diverse communication requirements while serving as a unified platform for various services and applications. Before 5G finally takes over, existing technologies such as LTE-Advanced and WLAN are gradually evolving to fit new needs. Furthermore, communication requirements of wearable devices can be fulfilled in part by existing technologies. For example, MU-MIMO, together with beamforming in 802.11ac, can achieve a throughput of over 1 Gbps [92]. For wearables requiring high data rates, the future evolution of WLAN (IEEE 802.11 family in particular) can act as an alternative solution. Perhaps most importantly, the hardware cost and power consumption in specifications like WLAN and Bluetooth are more suitable for wearable devices. This article focuses on challenges as well as enabling technologies in wearable communications. The main contributions are as follows:

1. We evaluate design challenges and requirements for wearable communications and present a communication architecture that reflects recent industrial/academia research directions.

2. A list of detailed techniques are selected that can help alleviate the challenges. We emphasized on MAC/PHY design and present our research results accordingly.

7.2 Challenges for Wearable Communications

7.2.1 Power Constraints

The relatively small size of most wearable devices poses a challenge when it comes to fitting a conventional battery inside. For most wearables, battery power tends to be

proportional to the device size. Hence today's consumer electronics design must reserve a vast amount of space for batteries alone. Some wearable devices also mount power consuming components like network chips, GPS, and continuous monitor sensors. For healthcare sensors, it is also important to keep the device unobtrusive to patients, especially for the implantable ones. The typical battery life should last at least several years. To make matters worse, wearable devices become fashion icons; hence, bulky and/or heavy design will inevitably flop in the market.

7.2.2 Variations on Communication Requirements

Wearable communication requirements vary depending on different use cases, in terms of differing data rate, latency, and reliability. On the one hand, traffic growth has historically been a key driving forces for new generation wireless systems. Since 2014, VR/AR technologies have turned into a clear reality and gigabit/s throughput has practically become a demanding feature in wearable device consumer markets. The roadmap of LTE and 5G have both proposed a gigabit/s experience in the near future. But the cost might be too high for massive connections from wearables, and some specifications like latency requirements cannot meet wearable demands.

7.2.3 Dense Deployment of Wearable Devices

Wearables can help users see, hear, sense, and even feel the world, making it very common for one user to require multiple devices, such as a fitness tracker to maintain a healthy lifestyle, a VR/AR helmet for gaming and exploring a richer experience, and a smart glasses for virtual assistance and navigation. Such a usage scenario may not cause problems in rural areas. Yet in areas with high population density, capacity and connectivity issues can become exaggerated and turn into performance barriers [93].

While most wearable communications occur locally by using WLAN, the contention based WLAN MAC protocol could limit the number of devices that can be supported. For example, even in ideal conditions, a typical Wi-Fi router can support a maximum of 30 to 50 connected devices. With hundreds of people in a large conference room, the number of

wearable devices could reach thousands. Not even a large deployment of hotspots could solve that communication problem, due to severe interference.

7.2.4 Health Concerns

A major concern regarding wearable communications is human biological safety under radio frequency (RF) exposure. The human body absorbs electromagnetic radiation, which causes thermal or non-thermal heat in the affected tissues. Guidelines on RF exposure normally apply specific absorption rate (SAR) as the metric for frequencies below 6 GHz. For mmWave, since the absorption is low and the primary energy remain in the surface layer of the skin, power density (PD) instead of SAR is more suitable for evaluating the health effects. However, PD cannot evaluate the effect of certain transmission characteristics such as reflection well. Therefore, temperature elevation of a direct contact area is proposed as the appropriate metric for mmWave RF exposure in [94]. Besides, some tissues like eyes are especially vulnerable to mmWave radiation-induced heating and requires more attention. It is necessary to continually update regulations based on new materials, frequencies, device types, and transmitted powers. In addition, manufacturers must be educated with the newest research/regulations to better mitigate consumer concerns and promote this new technology.

7.2.5 Security

Due to the computing and power limitations of wearable devices, collected data may need to be shared with other devices, edge nodes, or the cloud for further processing. The ever-growing desire to improve health and lifestyle remarkably promotes information sharing, such shared data will inevitably contain sensitive and private information, such as location, heart rate, emotion, and disease history. Thus, any leak of information could cause serious problems for individuals. The challenges here are multifold, including how to protect data so that unauthorized people will not have access, how to ensure data is securely shared between the device and the cloud, and how to make sure data is securely stored in the cloud.

7.3 Enabling Architecture for Wearable Communications

This article presents a wearable communication architecture that combines heterogeneous cloud radio access networks (H-CRAN) [95], cloud/fog computing, and software defined networks (SDN), as shown in Fig. 7.1. The H-CRAN architecture leverages macro base station (MBS), small base station (SBS), and remote radio header (RRH) to facilitate various user connections and performance needs. High power MBSs provide blanket coverage and seamless mobility, while low power SBSs and RRHs enable local coverage and fulfill high capacity requirements [27]. Wearables can utilize both MBSs and SBSs for data offloading, thereby saving energy and achieving faster computation speeds. Furthermore, MBSs and SBSs can connect to baseband unit (BBU) pools or to cloud servers directly, using backhaul connections. These BBU pools can help achieve globally optimized mobile association, interference management, and cooperation. When MBSs/SBSs are directly connected to cloud servers, extremely computation-intensive but less delay-stringent tasks can be offloaded to cloud servers for more powerful processing. To further leverage cloud RAN benefits, RRHs can be set up very close to end wearable devices. RRHs can be designed to mainly possess radio front functionalities while BBU pools handle the majority PHY/MAC layer processing. Low power RRHs can be largely deployed to provide various communication needs, such as low latency, low transmission power, and high capacity. In addition, RRHs can integrate several transmission technologies, including Bluetooth, WLAN, and visible light communication (VLC), to help provide backward compatibility and enrich use cases.

BBU pools are connected by data servers, in which user-specific data, such as preferences, locations, activities predictions, and QoS requirements, are stored. Data can be generated by cloud servers with SDN-controlled backhaul, wherein SDN controllers play an important role. To be specific, SDN controllers are aware of network condition and user demands, send instructions to BBU pools to guide network traffic forwarding. Furthermore, local SDN controllers in each sub-network can abstract physical devices to virtual ones. For example, depending on the application, network resources can be sliced to form virtual ac-

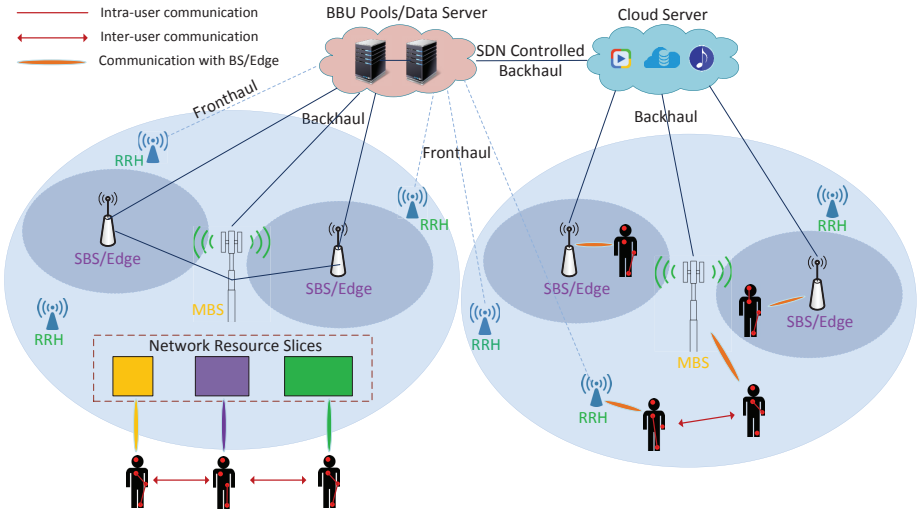

Fig. 7.1: An illustration of the wearable communication system architecture.

cess points (APs) to fit different needs [96]. Such network slicing could dynamically improve system performance.

Another important component is MEC. MEC is very important to wearable communications because it moves clouds locally to reduce transmission latency, backhaul loading, and the central node workload. One observation in wearable devices is that a significant portion of communications take place between various devices belonging to the same person. Therefore, D2D communications underlaying cellular network can transmit and process data locally, thereby reducing both latency and energy consumption. This cloud/edge architecture with D2D could greatly facilitate wearable communications. Communications between wearables and edge nodes can use different technologies, either on licensed bands or unlicensed bands.

1. Licensed Wearable Communications: Commercial cellular communications generally fall into this category. Within licensed communications, wearables can communicate with either edge nodes or BSs directly. Although QoS and mobility can be well supported through cellular networks, some potential drawbacks may exist. First, exclusive usage of spectrum elevates the expedient cost on carriers and ultimately increases customer cost. In addition, licensed communication chips tend to be more

Wearable Devices	Communication Requirements			Possible Specifications
	Latency	Capacity	Reliability	
AR/VR helmets	High	High	Low	mmWave Cellular &WLAN
Smartphones Tablets	Medium	Medium to High	Medium	LTE, Bluetooth, mmWave Cellular &WLAN
Medical sensors	High	Low	High	LTE, Bluetooth
Smart watch/glasses	Medium	Medium	Low	LTE, Bluetooth
Smart clothing/shoes	Low	Low	Low	Bluetooth, ZigBee

Table 7.1: Wearable communication requirements and possible solutions

complex and expensive. Further, licensed communication often consumes more power.

2. Unlicensed Wearable Communications: 2.4 GHz and 5 GHz WLAN, Bluetooth, and IEEE 802.15.4 are the most prevalent Sub-6 GHz unlicensed technologies used in today's wearable devices, with IEEE 802.11 ad (WiGig, 60 GHz WLAN) and visual light communications (VLC) still under research or under deployment. These specifications allow close proximity direct communications between two or more devices. Due in part to these factors, unlicensed communications enable cheaper and less complex devices, as well as a longer battery life, all of which are desirable for wearables. However, the maximum transmitted power constraint confines communications to a limited range at the unlicensed band. Furthermore, contention based access schemes of some techniques may impose certain limitations on the number of devices supported.

Table. 7.1 lists all the communication specifications and their potential use cases in wearable communications.

7.4 Enabling Transmission/Networking Technologies

7.4.1 Antenna Design

The form factor and power constraints on wearable devices impose additional requirements on antenna design. This is especially true for those devices operating on multiple modes, with transceivers designed to work in more than one protocol stack. As previously

mentioned, differing wireless technologies may use differing spectrum. Traditional cellular, Bluetooth, tri-band Wi-Fi (2.4 GHz, 5 GHz, and 60 GHz) and mmWave cellular are expected in the high-end wearable devices by the consumer market. Yet legacy antenna design requires the antenna size to be less than half of the wavelength, which can efficiently capture the radiated signal. For cellular systems, the frequency ranges from 800 MHz (GSM, low band 3G, and 4G) to around 2.5 GHz (high band 4G, Bluetooth, and 2.4 GHz Wi-Fi), antenna size varies from 18 cm to 5 cm. Today's innovative antenna design incorporates both engineering and industrial design aspects, that take advantage of the entire structure of a device. For smaller devices, patch antenna can be directly printed on the circuit board with a higher dielectric constant, thus reducing the required size at the cost of gain loss. Besides, [97] showed an innovative tri-band antenna design that can work in small-size wearables.

The mmWave frequency band not only leads to higher path loss, it is more susceptible to blockages, potential water vapor, and oxygen molecule absorption. However, a smaller wavelength at the mmWave band can benefit the antenna array design within a compact area. Hence, for mmWave communications, an antenna array with multiple antenna elements and directional beamforming could compensate for the downside of channel characteristics. In fact, as RF units drain a significant amount of battery energy when compared with other antenna components, the most advanced designs use fewer RF units, while still achieving a promising performance. The idea is to group several antenna elements into a single RF unit, thereby leading to a hybrid analog/digital antenna structure. Due to the sparse nature of multipath in mmWave signal, an improvement from pure digital beamforming is limited. And the relative simplicity of analog beamforming further motivates the hybrid analog and digital antenna structure. A prototype design made by Samsung Electronics has 32 antenna elements but only 4 RF units in a 6 cm $\times$ 3 cm area [98]. Such architecture can easily be applied to wearable devices such as AR helmets.

With regards to the BS, a large-scale antenna system (typically in the order of hundreds) can be mounted to form massive MIMO systems capable of serving more users with the

same time/frequency resources and provide a higher energy and spectral efficiency. Massive MIMO takes greater advantage of spatial diversity and/or multiplexing by supporting massive connectivities in a scalable way. The gain, however, comes primarily from an accurate knowledge of the CSI used for signal detection and precoding. How to obtain CSI using moderate to low overhead with pilot contamination in a massive MIMO system is an active yet challenging research. For wearable devices, due to their close proximity to the human body, antenna design must further consider the SAR. This regulation could impact the power level and antenna beam design.

7.4.2 PHY and MAC Technologies

The massive connectivity and high data rate requirements of wearable devices can be fulfilled, in part, by new radio access technologies (RATs) and MAC technologies. Emerging RATs, such as NOMA, benefit the system with both spectral efficiency and connectivity [19]. NOMA allows the same radio resources to be used by more than one wearable device at the same time, a contrast to current OMA technologies, such as orthogonal frequency-division multiple access (OFDMA) in 4G. The non-orthogonality can occur either in the power-domain (PD-NOMA) or the code domain (CD-NOMA). CD-NOMA utilizes different codes within the same resource to achieve multiplexing gain, while PD-NOMA assigns users with distinct power levels to maximize the performance.o In this chapter, we focus on PD-NOMA, denoted as NOMA for brevity. On the transmitter side, NOMA allocates more power to users with poor channel conditions, creating a power disparity which not only facilitates decoding, but also promotes system throughput and fairness. NOMA has the potential to achieve even higher spectral efficiency and connectivity at the cost of a more complex receiver structure, a problem which can be addressed by more advanced signal processing schemes and hardware design. To be specific, SIC is used to retrieve user messages by decoding the strongest signal first. SIC then subtracts the decoded signals and continue to decode the next strongest signal. This process stops once the intended received signal is decoded. Furthermore, D2D communication is considered a promising technology in 5G systems [99]. The D2D underlying cellular network allows direct communications between

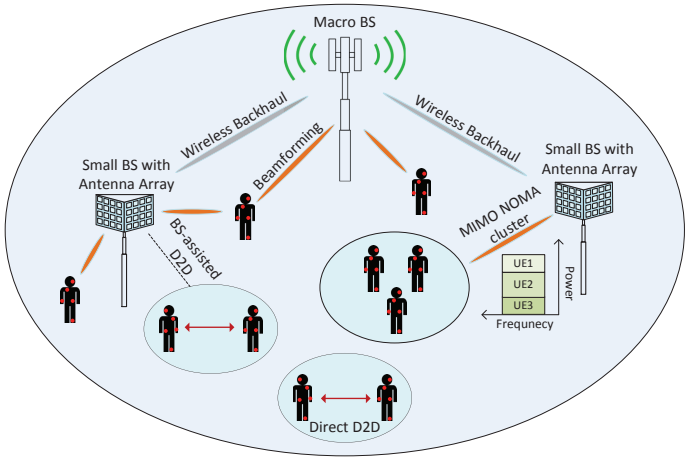

Fig. 7.2: A wearable communication system with MIMO, NOMA, and D2D PHY/MAC schemes

closely located users. Both D2D assisted cellular mode communications and direct D2D communications can utilize close proximity and frequency reuse gains so that higher energy and spectral efficiency can be achieved. A wearable communication system using NOMA, MIMO, and D2D PHY/MAC schemes is shown in Fig. 7.2.

Fig. 7.3 presents preliminary simulation results when multiuser-MIMO, NOMA, and D2D are applied to wearable communications. Specifically, consider a downlink wireless system with one edge node BS located in the center of a circle with a radius of R km. The BS has M antennas, whereas the cellular mode wearable devices (CWDs) and D pairs of D2D mode wearable devices (DWDs) have only one antenna each. Notice that a wearable can operate in both modes, depends on the connection requirement. CWD refers to the wearable that connect to cellular base stations for guaranteed service, while DWD is the mode for local connections. CWDs and DWDs are randomly deployed. The distance between each D2D transmitter and receiver is constant and denoted as R_d. The channel gain between the BS, CWDs, and DWDs consists of large-scale path loss and small-scale Rayleigh fading. To better utilize MIMO and NOMA, the edge node generates M beams and each beam supports K CWDs through NOMA. Thus in total $M \times K$ users can be supported on each radio resource unit. Precoding scheme and power allocation need to be optimized for maximizing the sum spectral efficiency of CWDs and DWDs. While the

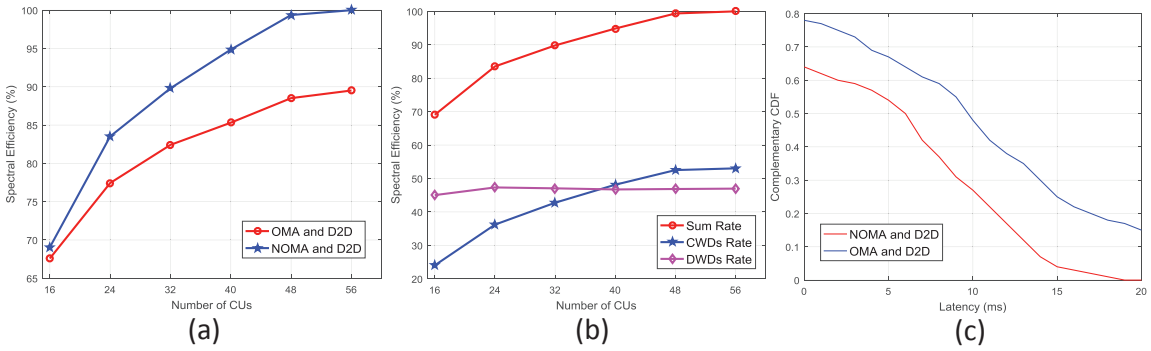

Fig. 7.3: Performance evaluation of a downlink system with MU-MIMO, NOMA and D2D. $R = 1$ km, $M = 4$, $K = 2$, $D = 2$, $R_d = 10$ m. Transmit powers of the edge node and DWDs are 10 and 2 Watts, respectively. (a), Sum rate of proposed NOMA+D2D with OMA+D2D. (b) The performance of CWDs and DWDs in NOMA+D2D scheme [19]. (c) CCDF performance w.r.t the latency.

joint optimization is difficult to achieve, we solve the problem in a heuristic way, wherein ZF precoding is determined first, followed by the NOMA power allocation is decided by applying KKT conditions in convex optimization. As a comparison, results from OMA are also presented, with only one CWD supported in each beam. All the results are expressed as the percentage for better illustration. Clearly, the proposed scheme reveals a better performance in terms of overall spectral efficiency, connectivity, and latency. And as the number of CWDs increases, the system can further benefit from multiuser diversity gain.

The second part is an advanced MAC protocol which coordinates transmission/processing within a device or between devices. Traditional methods dispatch transmissions according to varying protocol stacks within a device so that data going through BLE will not be sent to Wi-Fi. However, coordination can be made with a central unit (a dedicated low power always-on component) inside which takes a step further to send data to appropriate protocols, based on the availability, surrounding interference level, and demand for quality of experience (QoE). For example, emergency healthcare information can be dispatched to a cellular unit, resulting a fast response directly over the Internet. Voice calls can go through Wi-Fi if the cellular unit is unavailable. Besides, MAC plays a more important role in transmissions involving multiple devices. Smart transmission classifies data in terms of QoS requirements, which can help save battery life by forcing the RF unit to enter a sleep

mode that only activates to deliver critical information requiring low latency transmission. Further, since antennas can form narrow beams at the mmWave band, devices can support multiple transmissions simultaneously with limited co-channel interference [16]. MAC protocol needs to consider initial access aligning antenna to the high gain direction, thereby enabling beam-tracking in motions for seamless experience and an advanced sensing algorithm which senses the channel in a specified direction, rather than simply isotropically.

7.4.3 Cloud/Edge Computing

Cloud computing has brought significant changes to the Internet in the past few decades. Its centralized nature helps lower the expenditure cost while speeding up the deployment process. However, cloud computing alone cannot fulfill the demands of wearable communication. Cloud data centers are often located in remote regions, which may cause a long end-to-end latency, thereby impacting delay-sensitive applications. Since data are sent to the cloud for processing, concerns such as security and privacy possibly may arise as well. Yet current research now shifted to a combination of cloud and edge computing structure. Specifically, devices or nodes with storage, computing, and caching capabilities can be deployed in close proximity with wearable devices and act as middleware between cloud and local networks. These devices can be routers, small base stations, and even high-end wearable devices. In addition, advanced caching algorithms can offload popular contents from cloud to edge nodes either in real-time or offline. An illustration of the cloud/edge architecture is shown in Fig. 7.4. To better take advantage of varying spectrum, connections between edge nodes could utilize mmWave band which provides sufficient bandwidth for higher throughput [100], while the connection between devices belonging to the same person could use 2.4 GHz BLE and WLAN, or 5 GHz WLAN. By properly assigning spectrum, interference in the dense wearable networks can be reduced.

The advantages of this paradigm are multi-fold. Firstly, by providing certain computing capabilities via edge nodes or wearable devices, the transmission load on the backhaul can be greatly reduced. This benefit is prominent for applications such as online gaming, where 60 or even 120 frames need to be rendered per second. An alternative solution dictates that,

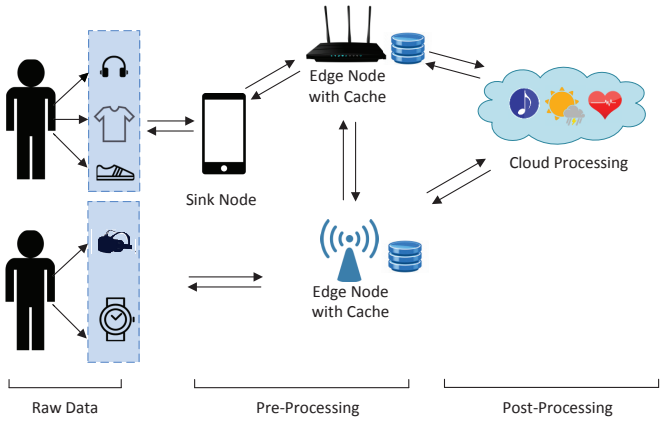

Fig. 7.4: Edge communication overview

servers only send parameters such as character's position, time-stamp, and property changes (few plain data) and allow the edge nodes to calculate and render visual images. Secondly, with the help of the large number of edge nodes deployed in 5G and big data analysis of user preferences, popular contents can be pre-fetched into connected edge devices, which are only one hop away from users. Thirdly, this scheme is more robust in terms of always-on connectivity, as well as privacy and security control. Lastly, cloud/edge computing enables a much more scalable architecture.

7.4.4 Energy Harvesting

The advancement of battery technologies lags behind its silicon counterparts. Nowadays, widely-used consumer device batteries are based on Lithium-ion. Researchers are working on improving battery energy density, finding new materials, and reducing charging time to deliver a better user experience. Meanwhile, energy efficiency has become a major concern in the network design. This problem can be alleviated by developing advanced energy harvesting techniques, which enable devices to harvest energy from the surrounding environment for both immediate and/or future usage by storing harvested energy in the battery unit. Such energy can come from solar power, ambient motion, the human body, background electromagnetic waves, etc. Solar power, for example, can be used to run outdoor wearables, such as edge nodes, watches, and smart clothes. Recent progress on solar

cell materials like perovskites, make the solar power harvest more flexible to integrate, and more efficient, not to mention cheaper. Ambient motion takes the advantage of mechanical movements by transferring them to electrical form. In general, direct force and inertial force on a proof mass are two main energy sources. Their principles, however, are similar. Since the generated energy of this technique is relatively low (a few microwatts, depending on specific activities), it is more promising for applications like foot-wear equipment and watches [101]. Furthermore, wearable devices can extract energy from the human body by capturing temperature differences between the body and the outside environment with a thermoelectric module. Even though the efficiency is limited, with only a few Celsius difference on average, its value has been proved by various commercialized products. In addition, studies have also reported the energy harvest from human body fluids. Lastly, energy harvest from electromagnetic waves is attracting more attention recently.

7.4.5 Advanced Security Solutions

Concerns with data security and privacy have increased, especially as users share more and more private data, including photos, locations, and activities. However, physical data collected from wearable devices such as medical conditions are sensitive and require extra protection. Typically, data goes through different phases, namely, data collection, transmission, and sharing. In data collection, biometric access is already widely used in high-end devices. Iris, face, and fingerprint recognition utilize specific user patterns to secure device access. Data can be further secured with schemes such as public key encryption. For some wearable applications, data needs to be shared to remote servers for analysis or diagnosis. Encryption requires collaboration with network architecture and transmission protocols. To be specific, we consider intra-wearable and inter-wearable communication scenarios. Intra-wearable communications occur between multiple devices carried by the same person while inter-wearable communications occur between multiple devices carried by different people. In the former case, biometric information such as inter-pulse interval (IPI) can be easily detected by multiple devices and can then be extracted for encryption and decryption key generation [102]. This extra protection at the protocol level makes

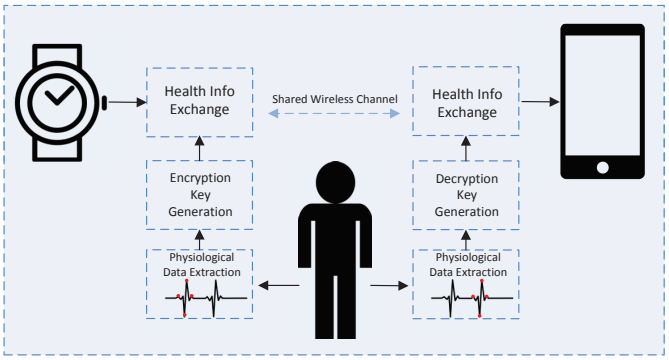

Fig. 7.5: An example of intra-wearable security solution using heart rate pattern

wearable communications more secure. A brief illustration is in Fig. 7.5. Inter-wearable communications, on the other hand, can leverage edge nodes and cloud servers. Specifically, public key cryptography can be made between wearables and servers for scalable considerations. For resource-constrained wearables, the impact of computational workload and power consumption on security should also be taken into consideration.

7.5 Chapter Conclusions

Recent explosive growth of wearable devices has spurred ever-increasing research interests in various fields, including communications. Yet such growth also presents paramount challenges in the same fields. In order to tackle these challenges, communications architecture and communication technologies are contemplating revolutionary changes. The multiple-layer communication architecture presented in this article combines D2D, C-RAN, and cloud/edge technologies into one to address stringent latency/power/computation constraints in wearable communications. Enabled by this multi-layer communications architecture, computation offloading to nearby devices, through D2D or to nearby edge nodes through cellular/other wireless technologies has been deemed one of the key techniques used to address fundamental wearable issues, such as limited battery, limited computing capability, critical latency on performance. Transmission technologies such as massive MIMO and NOMA applied wearable communications, can further significantly improve wearable communication spectral efficiency, power efficiency, and connectivity.

CHAPTER 8

Conclusions

8.1 Summary

In this dissertation, with the focus on achieving 5G ambitious goals on speed improvement, latency reduction, energy reservation, and massive connection, we propose different techniques to systematically make these goals feasible. In particular, we apply NOMA in various network settings, such as downlink MIMO, cooperative relay, LTE, and IoT networks. To tackle the error propagation in SIC process, we proposed a new model on residual interference. Additionally, we consider imperfect channel condition on NOMA and applied two general error estimation model in a downlink CR system. For the goal to reduce latency, we studied MEC, an emerging scheme which involved task offloading and local computing. Lastly, combining all the techniques above, we explored a communication architecture that can best suit wearable devices.

In part one, comprised of chapter 2 to 5, we mainly studied the performance of NOMA, which has the potential to fulfill the requirements on spectral efficiency, energy efficiency, and massive connection. Specifically, in chapter 2, NOMA was applied in downlink MIMO underlaid D2D networks. The contribution is with such a scheme, the base station can support more cellular users. Moreover, we designed two ZF-based beamforming, one is to compress interference to D2D users, the other is to suppress interference from other NOMA cellular groups. Both of them can improve the system total throughput. Furthermore, we reviewed current literatures on NOMA and revealed some strong assumptions on SIC process, then modelled decoding error from previous decoding as residual interference for current stage. The contribution is that we are one of the very first groups to propose and model this behavior. The challenge to solve the optimization problem involving the error propagation is the variables are coupled and we developed an iterative algorithm to

effectively get the precoding matrices. Next, we mainly explored the performance of NOMA in relay and IoT networks, respectively in chapter 4. Several schemes are investigated, our focus is to evaluate the performance with outage probability metric. Finally, in chapter 5, we considered channel uncertainty and the impact to NOMA, specifically, channel estimation error not only affected previous decoding (lead to error propagation), but also the current decoding. Two models are included, namely the bounded and Gaussian model.

In part two, as the main target to reduce delay, we studied joint offloading and local computing in MEC systems. The contribution is the new evaluation metric proposed, namely CE. We argued that CE is a more appropriate one that not just targeting on number of bit processed but also the energy consumed. CE can help find a trade-off between two processing schemes. The results also validated our algorithm and metrics.

In the last part, we applied the previous techniques into a potential application, wearable communication and its architecture. The challenges here is that wearable devices have diverse communication needs and current systems cannot fulfil all of them. We divided wearable device into 8 categories according to their communication requirements. Furthermore, the proposed architecture comprised NOMA and MEC in the physical layer, also we concluded that WLAN should play a very important role.

8.2 Future Works

8.2.1 Hardware Impairments for NOMA

To make NOMA more practical in real-world application, hardware impairments should be considered. That includes power amplifier saturation effect when transmission power is high, and quantization noise. In fact, there are few works on one-bit beamforming in NOMA, but the prior challenge has not been addressed.

8.2.2 CE in MEC Systems

In chapter 6, we only studied a simple scheme, the performance of CE under other communication settings have not been investigated. In particular, 1) MEC with cognitive

network. Devices may only be allowed to offload when the interference is low, not affecting primary users communication. We expect this scenario will be more practical. 2) MEC with NOMA. NOMA has the advantage on improving offloading rate, hence incorporating NOMA with MEC will be a good choice.

REFERENCES

[1] Cisco White paper, "Cisco visual networking index: global mobile data traffic forecast update, 2016-2021 White Paper," Mar. 2017.

[2] R. Q. Hu and Y. Qian, *Heterogeneous Cellular Networks*, John Wiley & Sons, Ltd., 2013.

[3] R. Q. Hu and Yi Qian, *Resource Management for Heterogeneous Networks in LTE Systems*, Springer, 2014.

[4] C. Lim, T. Yoo, B. Clerckx, B. Lee and B. Shim, " Recent trend of multiusesr MIMO in LTE-Advanced," *IEEE Commun. Mag*, vol. 52, no. 5, pp. 127-135, Mar. 2013.

[5] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro and K. Hugl, "Device-to-device communication as an underlay to LTE-advanced networks," *IEEE Commun. Mag*, vol 47, no. 12, pp. 42-49, Dec. 2009.

[6] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. Zhang, "What will 5G be?," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065-1082, June 2014.

[7] Q. Li, H. Niu, A. Papathanassiou, and G. Wu, "5G network capacity: key elements and technologies," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 7178, Mar. 2014.

[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 Sig. Proc. Lett.*, vol 21, no. 12, Dec. 2014.

[9] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for future radio access," in *Proc. IEEE TVC spring 2013*, Jun. 2013.

[10] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE PIMRC 2013*, pp. 611-515, Sep. 2013.

[11] Z. Zhang, H. Sun, and R. Q. Hu, "Downlink and uplink non-orthogonal multiple access in a dense wireless network," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2771-2784, Dec. 2017.

[12] A. R. Khan, M. Othman, S. A. Madani and S. U. Khan, "A survey of mobile cloud computing application models," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 393-413, Feb. 2014.

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

[14] P. Xu, Z. Ding, X. Dai and H. V. Poor, "NOMA: an information theoretic perspective," *arXiv preprint arXiv:1504.07751*, 2015.

[15] Y. Xu, H. Sun, R. Q. Hu, and Y. Qian, "Cooperative non-orthogonal multiple access in heterogeneous networks," in *Proc. IEEE GlOBECOMM 2015*, Dec. 2015.

[16] L. Wei, R. Q. Hu, T. He, and Y. Qian, "Device-to-device (D2D) communications underlaying MU-MIMO cellular networks," in *Proc. IEEE GlOBECOMM 2013*, pp. 4902-4907, Dec. 2013.

[17] B. Kim, S. Lim, H. Kim, S. Suh, J. Kwuun, S. Choi, C. Lee, S. Lee and D. Hong, "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *Proc. IEEE MILCOM 2013*, pp. 1278-1283, Nov. 2013.

[18] Q. H. Spencer, A. L. Swindlehurst and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Sig. Proc.*, vol. 52, no. 2, pp. 461-471, Feb. 2004.

[19] H. Sun, Y. Xu and R. Q. Hu, "A NOMA and MU-MIMO supported cellular network with underlaid D2D communications," *in Proc. IEEE VTC spring 2016*, May. 2016.

[20] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy efficiency of resource scheduling for non-orthogonal multiple access (NOMA) wireless network," *Proc. IEEE ICC 2016*, May. 2016.

[21] Q. Sun, S. Han, Z. Xu, S. Wang, C. I and Z. Pan, "Sun rate optimization for MIMO non-orthogonal multiple access systems," in *Proc. IEEE WCNC 2015*, pp. 747-752, Mar. 2015.

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

[23] M. Nonaka, A. Benjebbour, and K. Higuchi, "System-level throughput of NOMA using intra-beam superposition coding and SIC in MIMO downlink when channel estimation error exists," in *Proc. IEEE Int. Conf. on Commun. Systems*, pp. 202-206, Nov. 2014.

[24] J. G. Andrews and T. H. Meng, "Optimum power allocation for successive interference cancellation with imperfect channel estimation," *IEEE Trans. Wireless Commun.*, vol. 2, NO. 2, pp. 375-383, Mar. 2003.

[25] J. G. Andrews and T. H. Meng, "Performance of multicarrier CDMA with successive interference cancellation in a multipath fading channel," *IEEE Trans. Commun.*, vol. 51, no. 5, pp. 811-822, May. 2004.

[26] A. Goldsmith, S. A. Jafar and S. Vishwanath, "Capatity limits of MIMO channels," *IEEE J. Select. Areas Commun.*, vol 21, no. 5, pp. 684-702, Jun. 2003.

[27] R. Q. Hu and Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 94-101, May 2014.

[28] NTT DOCOMO, "Requirements, candidate solutions & technology roadmap for LTE Rel-12 onward," 3GPP RWS-120010, June 2012.

[29] N. Nonaka, Y. Kishiyama and K. Higuchi, "Non-orthogonal multiple access using intra-beam superposition coding and SIC in base station cooperative MIMO cellular downlink," in *Proc. IEEE VTC fall 2014*, Sept. 2014.

[30] J. Men and J. Ge, "Non-orthogonal multiple access for multiple-antenna relaying networks," *IEEE Commun. Lett.*, vol. 19, no. 10, pp. 1686-1689, Oct. 2015.

[31] J.-B. Kim and I.-H. Lee, "Non-orthogonal multiple access in coordinated direct and relay transmission," *IEEE Commun. Lett.*, vol. 19, no. 11, pp. 2037-2040, Nov. 2015.

[32] Z. Ding, H. Dai and H. V. Poor, "Relay selection for cooperative NOMA," *IEEE Wireless Commun. Lett.*, vol. 5, no. 4, pp. 416-419, Aug. 2016.

[33] 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 Fall 2016, invited paper*. Sept. 2016.

[34] A. K. Gupta and D. K. Nagar, *Matrix Variate Distributions*. London, U.K.: Champman & Hall, 2000.

[35] Y. Liu, Z. Ding, M. Ekashlan and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems with SWIPT," *2015 23rd European Signal Processing Conference (EUSIPCO)*, Nice France, Sept. 2015.

[36] E. Boshkovska, R. Morsi, D. W. K. Ng and R. Schober, "Power allocation and scheduling for SWIPT systems with non-linear energy harvesting model," *2016 IEEE International Conference on Communications (ICC)*, Kuala Lumpur, 2016, pp. 1-6.

[37] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Trans. Commun.*, (submitted) Available on-line at arxiv.org/abs/1607.06302.

[38] E. Boshkovska, D. W. K. Ng, N. Zlatanov and R. Schober, "Practical non-Linear energy harvesting model and resource allocation for SWIPT systems," in *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2082-2085, Dec. 2015.

[39] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products.* Academic press, 2014.

[40] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347-2381, Dec. 2017.

[41] D. Zhang, Y. Liu, Z. Ding, Z. Zhou, A. Nallanathan and T. Sato, "Performance analysis of non-regenerative massive-MIMO-NOMA relay systems for 5G," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4777-4790, Nov. 2017.

[42] X. Chen, Z. Zhang, C. Zhong, and D.W. K. Ng, "Exploitation multiple-antennas techniques for non-orthogonal multiple access," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2207-2220, Oct. 2017.

[43] Z. Ding, Z. Zhao, M. Peng and H. V. Poor, "On the spectral efficiency and security enhancements of NOMA assisted multicast-unicast streaming," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 3151-3163, July 2017.

[44] F. Zhou, N. C. Beaulieu, Z. Li, J. Si, and P. Qi, "Energy-efficient optimal power allocation for fading cognitive radio channels: Ergodic capacity, outage capacity and minimum-rate capacity," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2741-2755, Apr. 2016.

[45] A. Goldsmith, S. A. Jafar, I. Maric, and S. Srinivasa, "Breaking spectrum gridlock with cognitive radios: an information theoretic perspective," *Proc. IEEE*, vol. 97, no. 5, pp. 894-914, May 2009.

[46] Y. Liu, Z. Ding, M. Elkashlan, and J. Yuan, "Nonorthogonal multiple access in large-scale underlay cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10152-10157, Dec. 2016.

[47] L. Lv, J. Chen, Q. Li, and Z. Ding, "Design of cooperative non-orthogonal multicast cogntive radio multiple access for 5G systems: user scheduling and performance anaysis," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2641-2656, June 2017.

[48] L. Lv, Q. Ni, Z. Ding and J. Chen, "Application of non-orthogonal multiple access in cooperative spectrum-sharing networks over nakagami-m fading channels," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 5506-5511, June 2017.

[49] X. Huang, T. Han, and N. Ansari, "On green energy powered cognitive radio networks," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 827-842, Second Quarter, 2015.

[50] X. Lu, P. Wang, D. Niyato, D. I. Kim, and Z. Han, "Wireless networks with RF energy harvesting: A contemporary survey," *IEEE Commun. Surveys Tuts.*, vol. 17, pp. 757-789, Second Quarter, 2015.

[51] E. Boshkovska, D. W. K. Ng, N. Zlatanov, A. Koelpin, and R. Schober, "Robust resource allocation for MIMO wireless powered communication networks based on a non-linear EH model," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 1984-1999, May 2017.

[52] H. Sun, F. Zhou, and Z. Zhang, "Robust beamforming design in a NOMA cognitive radio network relying on SWIPT," in *Proc. IEEE ICC*, Kansas City, MO, USA, 2018.

[53] G. Pan et al., "On secrecy performance of MISO SWIPT systems with TAS and imperfect CSI," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3831-3843, Sept. 2016.

[54] Z. Chu, H. Xing, M. Johnston, and S. Le Goff, "Secrecy rate optimizations for a MISO secrecy channel with multiple multi-antenna eavesdroppers," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 283-297, Jan. 2016.

[55] D. W. K. Ng, E. S. Lo, and R. Schober, "Multi-objective resource allocation for secure communication in cognitive radio networks with wireless information and power transfer," *IEEE Trans. Veh. Technol.*, vol. 20, no. 2, pp. 328-331, Feb. 2016.

[56] P. V. Tuan and I. Koo, "Optimal multiuser MISO beamforming for power-splitting SWIPT cognitive radio networks," *IEEE Access*, vol. 5, pp. 14141-14153, Jul. 2017.

[57] Z. Hu, N. Wei, and Z. Zhang, "Optimal resource allocation for harvested energy maximization in wideband cognitive radio network with SWIPT, " *IEEE Access* vol. 5, 23383-23394, Aug. 2017.

[58] L. Mohjazi, I. Ahmed, S. Muhaidat, M. Dianati and M. Al-Qutayri, "Downlink beamforming for SWIPT multi-user MISO underlay cognitive radio networks," *IEEE Commun. Lett.*, vol. 21, no. 2, pp. 434-437, Feb. 2017.

[59] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938-953, Apr. 2016.

[60] N. T. Do, D. B. da Costa, T. Q. Duong, and B. An, "A BNBF user selection scheme for NOMA-based cooperative relaying systems with SWIPT," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 664-667, Mar. 2017.

[61] Z. Yang, Z. Ding, P. Zhi, and N. A. Dhahir, "The impact of power allocation on cooperative non-orthogonal multiple access networks with SWIPT," *IEEE Trans. Wireless Commun.*, to be published, 2017.

[62] P. D. Diamantoulakis, K. N. Pappi, Z. Ding, and G. K. Karagiannidis, "Wireless-powered communications with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8422-8436, Oct. 2016.

[63] P. D. Diamantoulakis, K. N. Pappi, G. K. Karagiannidis, H. Xing, and A. Nallanathan, "Joint downlink/uplink design for wireless powered networks with interference, " *IEEE Access*, vol. 5, pp. 1534-1547, 2017.

[64] Y. Xu, C. Shen, Z. Ding, X. Sun, S. Yan, G. Zhu, and Z. Zhong, "Joint beamforming and power-splitting control in downlink cooperative SWIPT NOMA systems," *IEEE Trans. Signal Process.*,vol 65, no. 18, pp. 4874-4886. Sept. 2017.

[65] F. Zhou, Z. Li, J. Cheng, Q. Li, and J. Si, "Robust AN-aided beamforming and power splitting design for secure MISO cognitive radio with SWIPT," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2450-2464, April 2017.

[66] B. Fang, Z. Qian, W. Zhang, and W. Shao, "AN-aided secrecy precoding for SWIPT in cognitive MIMO broadcast channels," *IEEE Commun. Lett.*, vol. 19, no. 9, pp. 1632-1635, Sept. 2015.

[67] C. Xu, Q. Zhang, Q. Li, Y. Tan, and J. Qin, "Robust transceiver design for wireless information and power transmission in underlay MIMO cognitive radio networks," *IEEE Commun. Lett.*, vol. 18, no. 9, pp. 1665-1668, Sept. 2014.

[68] F. Zhou, N. C. Beaulieu, J. Cheng, Z. Chu, and Y. Wang, "Robust max-min fairness resource allocation in sensing based wideband cognitive radio with SWIPT: Imperfect channel sensing," *IEEE Syst. Journal*, vol. PP, no. 99, pp. 1-12, Jun. 2017.

[69] T. A. Le Q.-T. Vien H. X. Nguyen D. W. K. Ng R. Schober "Robust chance-constrained optimization for power-efficient and secure SWIPT systems," *IEEE Trans. Green Commun. Netw.* vol. 1 no. 3 pp. 333-346 Sep. 2017.

[70] Y. Yuan and Z. Ding, "Outage constrained secrecy rate maximization design with SWIPT in MIMO-CR systems," *IEEE Trans. Veh. Technol.*, to be published, 2017.

[71] E. Boshkovska, D. W. K. Ng, L. Dai and R. Schober, "Power-Efficient and secure W-PCNs with hardware impairments and non-linear EH circuit," *IEEE Trans. Commun.*, to be published.

[72] E. Boshkovska, A. Keolpin, D. W. K. Ng, N. Zlatanov, and R. Schober, "Robust beamforming for SWIPT systems with non-linear energy harvesting model," in *Proc. IEEE Inter. Signal Processing Adances in Wireless Communications*, Edinburgh, UK, 2016.

[73] K. Xiong, B. Wang, and K. J. Ray Liu, "Rate-energy region of SWIPT for MIMO broadcasting under nonlinear energy harvesting model," *IEEE Trans. Wireless Commun.*, to be published, 2017.

[74] E. Boshkovska, N. Zlatanov, L. Dai, D. W. K. Ng, and R. Schober, "Secure SWIPT networks based on a non-linear energy harvesting model," in *Proc. IEEE WCNC 2017*, San Francisco, CA, USA, 2017.

[75] Y. Wang, Y. Wang, F. Zhou, Y. Wu, and H. Zhou, "Resource allocation in wireless powered cognitive radio networks based on a practical non-linear energy harvesting model," *IEEE Access*, to be published, 2017.

[76] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[77] K. Y. Wang, A. M. C. So, T. H. Chang, W. K. Ma, and C. Y. Chi, "Outage constrained robust transmit optimization for multiuser MISO downlinks: tractable approximations by conic optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5690-5705, Nov. 2014.

[78] I. Bechar, "A Bernstein-type inequality for stochastic processes of quadratic forms of Gaussian variables," [Online]. Avaliable: http://arxiv.org/abs/0909.3595.

[79] Y. Jong, "An efficient global optimization algorithm for nonlinear sum-of-ratios problem," [Online]. Available: http://www.optimization-online.org/DB_FILE/2012/08/3586.pdf

[80] E. Boshkovska, X. Chen, L. Dai, D. W. K. Ng, and R. Schober, "Max-min fair beamforming for SWIPT systems with non-linear EH model," [Online]. Available: https://arxiv.org/pdf/1705.05029.pdf

[81] F. Tang, Z. M. Fadlullah, N. Kato, F. Ono and R. Miura, "AC-POCA: Anti-coordination game based partially overlapping channels assignment in combined UAV

and D2D based networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1672-1683, Feb. 2018.

[82] J. Liu, H. Nishiyama, N. Kato, and J. Guo, "On the outage probability of device-to-device communication enabled multi-channel cellular networks: a RSS threshold-based perspective," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 163-175, Jan. 2016.

[83] T. G. Rodrigues, K. Suto, H. Nishiyama, N. Kato and K. Temma, "Cloudlets activation scheme for scalable mobile edge computing with transmission power control and virtual machine migration," *accepted by IEEE Trans. Comput.*, 2018.

[84] T. G. Rodrigues, K. Suto, H. Nishiyama and N. Kato, "Hybrid method for minimizing service delay in edge cloud computing through VM migration and transmission power control," *IEEE Trans. Comput.*, vol. 66, no. 5, pp. 810-819, May. 2017.

[85] S. Bi and Y. Zhang, "Computation rate maximization for wireless powered mobile-edge computing with binary computation offloading," [Online]. Available: https://arxiv.org/pdf/1708.08810.pdf

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

[87] F. Wang, J. Xu, X. Wang and S. Cui, "Joint offloading and computing optimization in wireless powered mobile-edge computing systems," submitted to *IEEE Trans. Wireless Commun.*, https://arxiv.org/abs/1702.00606.

[88] X. Cao, F. Wang, J. Xu, R. Zhang and S. Cui, "Joint computation and communication cooperation for mobile edge computing," Available: https://arxiv.org/pdf/1704.06777.pdf

[89] L. Yang, H. Zhang, M. Li, J. Guo and H. Ji, "Mobile edge computing empowered energy efficient task offloading in 5G," *accepted by IEEE Trans. Veh. Technol.*, 2018

[90] Q. Wu, W. Chen, D. W. Kwan Ng, J. Li and R. Schober, "User-centric energy efficiency maximization for wireless powered communications," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6898-6912, Oct. 2016.

[91] Y. Wang, M. Sheng, X. Wang, L. Wang and J. Li, "Mobile-edge computing: partial computation offloading using dynamic voltage scaling," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4268-4282, Oct. 2016.

[92] T. Baykas, C.-S. Sum, Z. Lan, and J. Wang, "IEEE 802.15.3c: The First IEEE Wireless Standard for Data Rates over 1 Gb/s," *IEEE Commun. Mag.*, vol. 49, no. 7, pp. 114-121, Jul. 2011.

[93] K. Venugopal and R. W. Heath, Jr., "Millimeter wave networked wearables in dense indoor environments," *IEEE Access*, vol. 4, pp. 1205-1221, Mar. 2016.

[94] T. Wu, T. S. Rappaport, and C. M. Collins, "Safe for generations to come: considerations of safety for millimeter waves in wireless communications," *IEEE Microw. Mag.*, vol. 16, no. 2, pp. 65-84, Mar. 2015.

[95] M. Peng, Y. Li, J. Jiang, J. Li and C. Wang, "Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies", *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126-135, Dec. 2014.

[96] K. Liang, L. Zhao, X. Chu and H. H. Chen, "An integrated architecture for software defined and virtualized radio access networks with fog computing," *IEEE Network*, vol. 31, no. 1, pp. 80-87, Jan.-Feb. 2017.

[97] M. N. Shakib, M. Moghavvemi and W. N. L. Binti Wan Mahadi, "Design of a tri-band off-body antenna for WBAN communication," *IEEE Antennas Wireless Propag. Lett.*, vol. 16, pp. 210-213, 2017.

[98] W. Hong, K. H. Baek, Y. Lee, Y. Kim and S. T. Ko, "Study and prototyping of practically large-scale mmWave antenna systems for 5G cellular devices," *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 63-69, Sept. 2014.

[99] L. Wei, R. Q. Hu, Y. Qian and G. Wu, "Enable device-to-device communications underlaying cellular networks: challenges and research aspects," *IEEE Commun. Mag.*, vol. 52, no. 6, pp. 90-96, June 2014.

[100] X. Ge, S. Tu, G. Mao, C. X. Wang and T. Han, "5G ultra-dense cellular networks", *IEEE Wireless Commun.* vol. 23, no. 1, pp. 72-79, Feb. 2016.

[101] D. E. Boyle, M. E. Kiziroglou, P. D. Mitcheson and E. M. Yeatman, "Energy provision and storage for pervasive computing," *IEEE Pervasive Comput.*, vol. 15, no. 4, pp. 28-35, Oct.-Dec. 2016.

[102] S. Wang, R. Bie, F. Zhao, N. Zhang, X. Cheng and H. A. Choi, "Security in wearable communications," *IEEE Network*, vol. 30, no. 5, pp. 61-67, Sept.-Oct 2016.

APPENDICES

APPENDIX A

Proof of Theorem 6

To prove the Theorem, we fist consider the KKT conditions of $\mathbf{P}_3$. Specifically, with some simple algebraic manipulation, (5.12) can be rewritten as

$$\begin{bmatrix} \alpha_{i,k}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & t_{i,k} \end{bmatrix} + \begin{bmatrix} \mathbf{I} \\ \widehat{\mathbf{h}}_i^{\dagger} \end{bmatrix} \mathbf{C}_k \begin{bmatrix} \mathbf{I} & \widehat{\mathbf{h}}_i \end{bmatrix} + \begin{bmatrix} -\gamma_{k,\min}\sum_{j=1}^{k-1}\mathbf{W}_j & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix} \succeq \mathbf{0}, \qquad \forall k \in \mathcal{K},\ i = \{k, k+1, \ldots, K\}, \tag{A.1}$$

where $t_{i,k} = -\alpha_{i,k}\varphi_k^2 - \gamma_{k,\min}\big(\sigma_{k,S}^2 + \frac{\sigma_D^2}{(1-\rho)}\big)$.

Similarly, (5.14) and 5.16 can be rewritten as

$$\begin{bmatrix} \beta_n\mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\beta_n\psi_n^2 + P_{n,p} \end{bmatrix} - \begin{bmatrix} \mathbf{I} \\ \widehat{\mathbf{g}}_n^{\dagger} \end{bmatrix} \mathbf{\Sigma} \begin{bmatrix} \mathbf{I} & \widehat{\mathbf{g}}_n \end{bmatrix} \succeq \mathbf{0},\ \forall n \in \mathcal{N}, \tag{A.2}$$

and

$$\begin{bmatrix} \theta_k\mathbf{I} & \mathbf{0} \\ \mathbf{0} & m_k \end{bmatrix} + \begin{bmatrix} \mathbf{I} \\ \widehat{\mathbf{h}}_k^{\dagger} \end{bmatrix} \mathbf{\Sigma} \begin{bmatrix} \mathbf{I} & \widehat{\mathbf{h}}_k \end{bmatrix} \succeq \mathbf{0},\ \forall k \in \mathcal{K}, \tag{A.3}$$

respectively, where $m_k = -\theta_k\varphi_k^2 + \sigma_{k,S}^2 - \frac{\tau_k}{\rho}$.

For notational simplicity, we let $\mathbf{X}_i = \begin{bmatrix} \mathbf{I} & \widehat{\mathbf{h}}_i \end{bmatrix}$ and $\mathbf{Y}_n = \begin{bmatrix} \mathbf{I} & \widehat{\mathbf{g}}_n \end{bmatrix}$. Also, denote $\mathbf{A}_{i,k} \in \mathbb{C}_+^{(M+1)\times(M+1)}$, $\mathbf{B}_k \in \mathbb{C}_+^{(M+1)\times(M+1)}$, $\mathbf{D}_n \in \mathbb{C}_+^{(M+1)\times(M+1)}$, $z \in \mathbb{R}_+$, and $\mathbf{E}_k \in$

$\mathbb{C}_{+}^{(M)\times(M)}$ as the KKT multiplier. Then the Lagrange dual function $\mathcal{L}$ can be expressed as

$$\mathcal{L}(\mathbf{W}_k, \mathbf{V}, \mathbf{A}_{i,k}, \mathbf{B}_k, \mathbf{D}_n, z, \kappa) = \mathrm{Tr}(\mathbf{\Sigma}) - \sum_{i,k} \mathrm{Tr}(\mathbf{A}_{i,k}\mathbf{X}_i^\dagger \mathbf{C}_k \mathbf{X}_i) - \sum_{i,k} \mathrm{Tr}(\mathbf{A}_{i,k}\mathbf{M}_k) \tag{A.4}$$
$$+ \sum_n \mathrm{Tr}(\mathbf{D}_n \mathbf{Y}_n^\dagger \mathbf{\Sigma} \mathbf{Y}_n) - \sum_k \mathrm{Tr}\big(\mathbf{B}_k \mathbf{X}_k^\dagger \mathbf{\Sigma} \mathbf{X}_k\big) + z\big(\mathrm{Tr}(\mathbf{\Sigma}) - P_B\big) - \sum_k \mathrm{Tr}(\mathbf{E}_k \mathbf{W}_k) + \kappa,$$

where $\mathbf{M}_k = \begin{bmatrix} -\gamma_{k,\min} \sum_{j=1}^{k-1} \mathbf{W}_j & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix}$ and κ are the terms irrelevant of $\mathbf{W}_k$. Taking the partial derivative of the dual function regarding $\mathbf{W}_k$, we have

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_k} = \mathbf{I} - \sum_i \mathbf{X}_i \mathbf{A}_{i,k} \mathbf{X}_i^\dagger + \gamma_{k,\min} \sum_i \sum_{j=1}^{k-1} \mathbf{X}_i \mathbf{A}_{i,j} \mathbf{X}_i^\dagger + \sum_i \gamma_{k,\min} \sum_{j=k+1}^{K} \mathbf{A}_{i,j} \tag{A.5}$$
$$+ \sum_n \mathbf{Y}_n \mathbf{D}_n \mathbf{Y}_n^\dagger - \sum_k \mathbf{X}_k \mathbf{B}_k \mathbf{X}_k^\dagger + z\mathbf{I} - \mathbf{E}_k = \mathbf{0}.$$

In addition, the dual problem needs to satisfy the completeness slackness

$$\Big(\begin{bmatrix} \alpha_{i,k}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & t_{i,k} \end{bmatrix} + \mathbf{X}_i^\dagger \mathbf{C}_k \mathbf{X}_i + \mathbf{M}_k \Big) \mathbf{A}_{i,k} = \mathbf{0}, \tag{A.6a}$$
$$\mathbf{E}_k \mathbf{W}_k = \mathbf{0}, \forall k \in \mathcal{K}, i = \{k+1, \ldots, K\}, \forall n \in \mathcal{N}. \tag{A.6b}$$

Right multiplying $\mathbf{W}_k$ with (A.5), and substituting (A.6b), we can get

$$\Big(\sum_i \mathbf{X}_i \mathbf{A}_{i,k} \mathbf{X}_i^\dagger + \sum_k \mathbf{X}_k \mathbf{B}_k \mathbf{X}_k^\dagger\Big) \mathbf{W}_k = \big[(1+z)\mathbf{I} + \gamma_{k,\min} \sum_i \sum_{j=1}^{k-1} \mathbf{X}_i \mathbf{A}_{i,j} \mathbf{X}_i^\dagger \tag{A.7}$$
$$+ \gamma_{k,\min} \sum_i \sum_{j=k+1}^{K} \mathbf{A}_{i,j} + \sum_n \mathbf{Y}_n \mathbf{D}_n \mathbf{Y}_n^\dagger \big] \mathbf{W}_k.$$

Since all the KKT multipliers are positive numbers or positive semidefinite matrix, we can easily verify $\big\{(1+z)\mathbf{I} + \gamma_{k,\min} \sum_i \sum_{j=1}^{k-1} \mathbf{X}_i \mathbf{A}_{i,j} \mathbf{X}_i^\dagger + \gamma_{k,\min} \sum_i \sum_{j=k+1}^{K} \mathbf{A}_{i,j} + \sum_n \mathbf{Y}_n \mathbf{D}_n \mathbf{Y}_n^\dagger \big\} \succeq \mathbf{0}$. Thus it is non-singular. Left multiplying a non-singular matrix with $\mathbf{W}_k$ does not change

the rank of $\mathbf{W}_k$. Therefore, we have

$$\begin{aligned} \text{Rank}(\mathbf{W}_k) &= \text{Rank}\big((\sum_i \mathbf{X}_i\mathbf{A}_{i,k}\mathbf{X}_i^\dagger + \sum_k \mathbf{X}_k\mathbf{B}_k\mathbf{X}_k^\dagger)\mathbf{W}_k\big) \\ &= \min\big\{\text{Rank}\big(\sum_i \mathbf{X}_i\mathbf{A}_{i,k}\mathbf{X}_i^\dagger + \sum_k \mathbf{X}_k\mathbf{B}_k\mathbf{X}_k^\dagger\big), \text{Rank}(\mathbf{W}_k)\big\}. \end{aligned} \tag{A.8}$$

Next, we show the rank of $(\sum_i \mathbf{X}_i\mathbf{A}_{i,k}\mathbf{X}_i^\dagger)$ is less than or equal to 2. By summing (A.6a) in terms of index i, then left-multiplying $\begin{bmatrix} \mathbf{I}_M & \mathbf{0} \end{bmatrix}$ and right-multiplying $\mathbf{X}_i^\dagger$, we have

$$\begin{aligned} &\sum_i \alpha_{i,k}\mathbf{X}_i\mathbf{A}_{i,k}\mathbf{X}_i^\dagger - \sum_i \alpha_{i,k}\begin{bmatrix} \mathbf{0}_M & \mathbf{h}_i \end{bmatrix}\mathbf{A}_{i,k}\mathbf{X}_i^\dagger + \sum_i \mathbf{C}_k\mathbf{X}_i\mathbf{A}_{i,k}\mathbf{X}_i^\dagger \\ &+\sum_i(-\gamma_{k,\min}\sum_{j=1}^{k-1}\mathbf{W}_j)\mathbf{X}_i\mathbf{A}_{i,k}\mathbf{X}_i^\dagger - \sum_i(-\gamma_{k,\min}\sum_{j=1}^{k-1}\mathbf{W}_j)\begin{bmatrix} \mathbf{0}_M & \mathbf{h}_i \end{bmatrix}\mathbf{A}_{i,k}\mathbf{X}_i^\dagger = \mathbf{0}. \end{aligned} \tag{A.9}$$

After a simple transformation, we have

$$\sum_i(\alpha_{i,k}\mathbf{I}+\mathbf{C}_k-\gamma_{k,\min}\sum_{j=1}^{k-1}\mathbf{W}_j)\mathbf{X}_i\mathbf{A}_{i,k}\mathbf{X}_i^\dagger=\sum_i(\alpha_{i,k}\mathbf{I}-\gamma_{k,\min}\sum_{j=1}^{k-1}\mathbf{W}_j)\begin{bmatrix} \mathbf{0}_M & \mathbf{h}_i \end{bmatrix}\mathbf{A}_{i,k}\mathbf{X}_i^\dagger \tag{A.10}$$

From the fact that (5.12) is a positive semidefinite matrix, $\big(\alpha_{i,k}\mathbf{I} + \mathbf{C}_k - \gamma_{k,\min}\sum_{j=1}^{k-1}\mathbf{W}_j\big)$ would be a non-singular matrix, thus the rank of the left term of the above equation is the same as $\sum_i \mathbf{X}_i\mathbf{A}_{i,k}\mathbf{X}_i^\dagger$. Also, it is easy to verify that the right term has a rank 1.

Similarly, we can prove that $\text{Rank}(\sum_n \mathbf{Y}_n\mathbf{D}_n\mathbf{Y}_n^\dagger) = 1$. Therefore, the following equation holds.

$$\begin{aligned} \text{Rank}\big(\sum_i \mathbf{X}_i\mathbf{A}_{i,k}\mathbf{X}_i^\dagger + \sum_k \mathbf{X}_k\mathbf{B}_k\mathbf{X}_k^\dagger\big) &\le \text{Rank}(\sum_i \mathbf{X}_i\mathbf{A}_{i,k}\mathbf{X}_i^\dagger) + \text{Rank}(\sum_n \mathbf{Y}_n\mathbf{D}_n\mathbf{Y}_n^\dagger) \\ &= 2, \end{aligned} \tag{A.11}$$

which proves the theorem.

CURRICULUM VITAE

Haijian Sun

Haijian Sun received the B.S. and M.S. degrees in electrical engineering from Xidian University, Xian, China, in 2011 and 2014, respectively. Since 2014, He has been a Ph.D. candidate with the Department of Electrical and Computer Engineering, Utah State University, Logan, Utah, under the supervision of Prof. Rose Qingyang Hu. His research interests include massive MIMO, non-orthogonal multiple access, SWIPT, wearable and IoT communications, machine learning applications on vehicular and drone communications, and 5G PHY. His industry experience includes work at Mitsubishi Electric Research Laboratories during the summer of 2018. He is the recipient of "PhD Scholar of the Year" in the College of Engineering, Utah State University, 2019.

Published Journal Articles

- H. Sun, F. Zhou, and R. Q. Hu, "Joint Offloading and Computation Energy Efficiency Maximization in a Mobile Edge Computing System", *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 3052-3056, Mar. 2019
- H. Sun, F. Zhou, R. Q. Hu, and L. Hanzo, "Robust Beamforming Design in a NOMA Cognitive Radio Network Relying on SWIPT", *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 1, pp. 142-155, Jan. 2019.
- H. Sun, Z. Zhang, R. Q. Hu, and Y. Qian, "Wearable Communications in 5G: Challenges and Enabling Technologies", *IEEE Vehicular Technology Magazine*, vol. 13, no. 3, pp. 100-109, Sept. 2018.
- F. Zhou, Z. Chu, H. Sun, R. Q. Hu, and L. Hanzo, "Artificial Noise Aided Secure Cognitive Beamforming for Cooperative MISO-NOMA Using SWIPT", *IEEE Journal on Selected Areas in Communications.*, vol 36, no. 4, pp. 918-931, Apr. 2018.

- Z. Zhang, H. Sun, and R. Q. Hu, "Downlink and Uplink Non-Orthogonal Multiple Access in a Dense Wireless Network", *IEEE Journal on Selected Areas in Communications,* vol. 35, no. 12, pp. 2771-2784, Dec. 2017.

Published Conference Papers

- H. Sun, P. Wang, M. Pajovic, T. Koike-Akino, P. Orlik, A. Taira, and K. Nakagawa, "mmWave Localization with 28-GHz Channel Measurement: A Preliminary Field Study", *submitted, in Proc. IEEE ICC 2019.*

- F. Zhou, H. Sun, Z. Chu, and R. Q. Hu, "Computation Efficiency Maximization for Wireless-Powered Mobile Edge Computing", *to appear, in Proc. IEEE Globecom 2019.*

- H. Sun, F. Zhou, and Z. Zhang, "Robust Beamforming Design in a NOMA Cognitive Radio Network Relying on SWIPT", in *Proc. IEEE ICC 2018.*

- F. Zhou, Z. Chu, H. Sun, and V. C. M. Leung, "Resource Allocation for Secure MISO-NOMA Cognitive Radios Relying on SWIPT", in *Proc. IEEE ICC 2018.*

- F. Zhou, Y. Wu, H. Sun, and Z. Chu, "UAV-Enabled Mobile Edge Computing: Offloading Optimization and Trajectory Design", in *Proc. IEEE ICC 2018.*

- H. Sun, Q. Wang, R. Q. Hu, "Outage Probability in a NOMA Relay System, " in *Proc. IEEE WCNC,* San Francisco, May 2017.

- H. Sun, Q. Wang, S. Ahmed and R. Q. Hu, "Non-Orthogonal Multiple Access in a mmWave Based IoT Wireless System with SWIPT," invited paper in *Proc. VTC Spring*, Sydney, NSW, 2017.

- Y. Xu, H. Sun, R. Q. Hu, "Hybrid MU-MIMO and Non-orthogonal Multiple Access Design in Wireless Heterogeneous Networks," in *Proc. EUSIPCO 2016,* Budapest, Hungary, Sept. 2016.

- Z. Zhang, H. Sun, R. Q. Hu, "Stochastic Geometry Based Performance Study on 5G Non-Orthogonal Multiple Access Scheme", in *Proc. IEEE GLOBECOM 2016*, Washington DC, Dec. 2016.

- H. Sun, B. Xie, R. Q. Hu and G. Wu, "Non-orthogonal multiple access with SIC Error Propagation in Downlink Wireless MIMO Networks," invited paper, in *Proc. IEEE VTC Fall,* Montreal, Canada, Sept. 2016.

- H. Sun, Y. Xu and R. Q. Hu, "A NOMA and MU-MIMO Supported Cellular Network with Underlaid D2D Communications", in *Proc. IEEE VTC Spring*, May. 2016.

- Y. Xu, H. Sun, R. Q. Hu, and Y. Qian, "Cooperative Nonorthogonal Multiple Access in Heterogeneous Networks," in *Proc. IEEE GlOBECOM*, San Diego, Dec. 2015.