Imperial College London
Department of Electrical and Electronic Engineering

Uplink Rate-Splitting Multiple Access for 6G

Jiawei Xu

Supervised by Prof. Bruno Clerckx

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy of Imperial College London and
the Diploma of Imperial College London, 2024

Declaration of Originality

I hereby declare that the content of this thesis is my own research work. The main parts of this thesis have been published in related journals. Where other sources of information have been used, they have been acknowledged.

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

Rate-Splitting Multiple Access (RSMA) is a pioneering framework that optimises non-orthogonal transmission, Multiple Access, and interference management in next-generation wireless networks. This thesis assesses uplink RSMA performance for both Single-Input Single-Output (SISO) and Multiple-Input Multiple-Output (MIMO) channels in the Finite Blocklength (FBL) regime, and designs the Physical (PHY)-layer architecture for uplink MIMO RSMA. Additionally, it explores the implementation of RSMA in Mobile Edge Computing (MEC) and evaluates various scenarios.

First, the performance of uplink SISO RSMA is investigated under FBL constraints with and without time-sharing. Sum-throughput and rate region maximisation problems by optimising power and rate allocations are formulated and solved using the Successive Convex Approximation (SCA)-based algorithm. The impact of blocklength and transmission rate on the throughput and error probability performance of RSMA are investigated. According to simulation results, RSMA is demonstrated to be promising for FBL communications.

Next, the research scope is extended from uplink SISO RSMA in the FBL regime to uplink MIMO RSMA. An uplink MIMO RSMA framework is proposed and a problem optimising both precoders and combiners with Max-Min Fairness (MMF) metric is investigated. The Alternating Optimisation (AO) is applied to decompose the problem into two subproblems due to the high coupling between precoders and combiners. The PHY-layer uplink MIMO RSMA architecture is designed. The effectiveness and robustness of the proposed uplink MIMO RSMA schemes are demonstrated through numerical results.

Last, the application of RSMA with the framework of MEC in short-packet communication is examined. A successful computation probability maximisation problem is formulated by optimising the offloading factor, power allocation and task-splitting factor. The optimisation problem is solved by SCA-based algorithm. The results indicate that RSMA is reliable for communication in MEC under FBL constraints.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Bruno Clerckx. Your guidance and mentorship not only provided me with invaluable academic direction but also support on my journey. Your expertise and advice propelled me forward in my research, and I will forever be grateful.

I also want to thank my colleagues and friends. Throughout our studies, we faced challenges together, worked hard, and supported each other. Your friendship and camaraderie have been a precious asset on my journey. Your understanding and support during my academic life enabled me to maintain balance and happiness. Your companionship has brought warmth and joy to my life.

Most importantly, I am indebted to my parents. Thank you for your unwavering support and encouragement, which have been the driving force behind my pursuit of my dreams.

Lastly, I would like to thank myself for not giving up. Throughout the process of the whole PhD, I encountered various difficulties and challenges, but I survived. It is my resilience and determination that have brought me to this point today, and I will continue to strive for self-improvement.

Jiawei Xu

London, May 2024.

Abbreviations

1G First Generation

2G Second Generation

3G Third Generation

4G Fourth Generation

5G Fifth Generation

6G Sixth Generation

AMC Adaptive Modulation and Coding

AO Alternating Optimisation

AoI Age of Information

AR Augmented Reality

AWGN Additive White Gaussian Noise

DF Decode-and-Forward

B5G Beyond Fifth-Generation

BC Broadcast Channel

BCD Block Coordination Descent

BLER Block Error Rate

BS Base Station

CDMA Code Division Multiple Access

CPU Central Processing Unit

CSIR Channel State Information at Receiver

CSIT Channel State Information at Transmitter

CSI Channel State Information

DoF Degree of Freedom

DPC Dirty Paper Coding

EE Energy Efficiency

eMBB enhanced Mobile Broadband

FBL Finite Blocklength

FDMA Frequency Division Multiple Access

FEC Forward Error Correction

HRS Hierarchical RS

GRS Generalised RS

i.i.d independent and identically distributed

IFBL Infinite Blocklength

IoT Internet of Things

LHS Light-Hand Side

LLR Log-Likelihood Ratios

LLS Link Level Simulations

MA Multiple Access

MCC Mobile Cloud Computing

MCS Modulation and Coding Scheme

MEC Mobile Edge Computing

ML Machine Learning

MU Multiple User

MU-LP Multi-User Linear Precoding

MAC Multiple Access Channel

MIMO Multiple-Input Multiple-Output

MISO Multiple-Input Single-Output

MMF Max-Min Fairness

MMSE Minimum Mean Square Error

mMTC massive Machine Type Communications

NOMA Non-Orthogonal Multiple Access

OFDMA Orthogonal Frequency Division Multiple Access

OMA Orthogonal Multiple Access

PHY Physical

QoS Quality-of-Service

RAN Radio Access Network

RF Radio Frequency

RIS Reconfigurable Intelligent Surface

RHS Right-Hand Side

RSMA Rate Splitting Multiple Access

RS Rate-Splitting

SCA Successive Convex Approximation

SCL Successive Cancellation List

SCP Successful Computation Probability

SC Superposition Coding

SC-SIC Superposition Coding-Successive Interference Cancellation

SD Soft Decision

SDMA Space Division Multiple Access

SE Spectral Efficiency

SIC Successive Interference Cancellation

SIMO Single-Input Multiple-Output

SINR Signal-to-Interference-plus-Noise-Ratio

SISO Single-Input Single-Output

SNR Signal-to-Noise Ratio

SOCP Second Order Cone Programming

TDMA Time Division Multiple Access

TDD Time Division Depulx

TIN Treat-Interference-as-Noise

URLLC Ultra-Reliable and Low-Latency Communications

VP Vector Perturbation

WMMSE Weighted Minimum-Mean Square Error

w.r.t with respect to

Notation

Objects

a	Scalar
$\mathbf{a}$	Column vector
$\mathbf{A}$	Matrix
$\mathbf{I}_N$	$N \times N$ Identity matrix
$\mathcal{A}$	Finite set
$\mathbb{1}$	Indicator function

Sets

$\mathbb{N}$	Natural numbers
$\mathbb{R}$	Real numbers
$\mathbb{C}$	Complex numbers

Operations

$(\cdot)^{\mathsf{H}}$	Hermitian (conjugate transpose)

$|\cdot|$ Magnitude of a complex number

$\|\cdot\|$ Euclidean norm of a vector

$\exp(\cdot)$ Element by element exponential of a vector

$\mathrm{tr}(\cdot)$ Trace of a square matrix

Superscript

$(\cdot)^{[n]}$ n-th iterated value

Contents

Declaration of Originality **1**

Copyright Declaration **3**

Abstract **5**

Acknowledgements **7**

Abbreviations **9**

Notation **13**

1 Introduction **25**

1.1 A New Multiple Access Technology: Rate-Splitting Multiple Access 26

1.2 Motivation 32

1.3 Organisation and Contributions 33

1.4 Publications 35

2 Preliminaries **36**

2.1 Fundamentals of RSMA 36

2.1.1 Downlink RSMA 36

2.1.2 Uplink RSMA 40

2.2 Finite Blocklength Regime 45

2.3 Mobile Edge Computing 50

3 Uplink SISO RSMA in Finite Blocklength Regime **55**

3.1 Introduction . . . 56
3.2 System Model . . . 57
3.2.1 RSMA . . . 57
3.2.2 NOMA . . . 59
3.3 Error Probability Analysis . . . 60
3.3.1 RSMA . . . 60
3.3.2 NOMA . . . 61
3.4 Problem Formulation and Algorithm . . . 62
3.4.1 Sum-throughput Maximisation . . . 62
3.4.2 Rate-region Maximisation . . . 67
3.5 Numerical Results and Discussion . . . 70
3.5.1 Sum-throughput . . . 70
3.5.2 Rate Region . . . 72
3.6 Summary . . . 74

4 Uplink MIMO RSMA in Finite Blocklength Regime . . . 75
4.1 Introduction . . . 76
4.2 System Model . . . 77
4.2.1 MIMO RSMA . . . 78
4.2.2 MIMO NOMA . . . 82
4.3 Problem Formulation and Optimisation . . . 84
4.3.1 Problem Formulation . . . 84
4.3.2 Precoders Optimisation . . . 87
4.3.3 Combiners Optimisation . . . 88
4.3.4 Proposed Algorithm, Convergence and Complexity . . . 89
4.4 Physical Layer Design . . . 91
4.5 Numerical Results and Disucssion . . . 94
4.5.1 MMF Performance . . . 94

4.5.2 Link-Level Simulation Results 99
4.6 Summary 102

5 RSMA with Mobile Edge Computing 103
5.1 Introduction 104
5.2 System Model 105
5.2.1 Tasks Offloading Phase 107
5.2.2 Tasks Computation Phase 110
5.3 Successful Computation Probability Analysis 111
5.3.1 Successful Offloading Probability P_s^o 111
5.3.2 Successful Execution Probability P_s^e 112
5.4 Problem Formulation and Algorithm 113
5.4.1 Offloading factor Optimisation 115
5.4.2 Transmit Power Optimisation 116
5.4.3 Task-Splitting Factor Optimisation 119
5.5 Numerical Results and Discussion 121
5.5.1 The SCP versus Task Size 122
5.5.2 The SCP versus Blocklength 124
5.5.3 The SCP versus Transmit SNR 126
5.6 Summary 128

6 Conclusions 129
6.1 Summary of Thesis Achievements 129
6.2 Future Work 132

References 133

List of Tables

2.1 Requirements of eMBB, URLLC and mMTC applications [1, 2] . . . 45

2.2 The importance of key capabilities in different usage scenarios [3] . 46

List of Figures

2.1 Transceiver architecture of K-user downlink RSMA. Message splitter is implemented by demultiplexer. 37

2.2 Capacity region of the two-user uplink AWGN channel. The two corner points, A and B, correspond to the maximum rates achieved by two Successive Interference Cancellation (SIC) decoding orders: A is achieved by decoding user-1 first and B is achieved by decoding user-2 first. P_1 and P_2 is the transmit power for user-1 and user-2, respectively. σ_n^2 is the noise variance at the receiver. 41

2.3 Transceiver architecture of K-user uplink RSMA. 42

2.4 Main computation components in a face recognition application [4]. 51

2.5 Main computation components in an AR application [5]. 52

3.1 Uplink SISO RSMA for the two-user system. 58

3.2 Throughput of Non-Orthogonal Multiple Access (NOMA) and Rate Splitting Multiple Access (RSMA) without time-sharing 71

3.3 Throughput of NOMA and RSMA with time-sharing. 72

3.4 Rate region of two users with $\epsilon = 10^{-3}$. 73

4.1 Transmission Model for uplink Multiple User (MU)-Multiple-Input Multiple-Output (MIMO) adopting RSMA at the user side. Users are classified into splitting users (depicted in orange block) and non-splitting users (depicted in yellow block), with each user in the former group splitting their respective messages to experience different decoding priorities at the Base Station (BS), leveraging the benefit of rate-splitting. 78

4.2 The schematic illustration of the two-user rate-region. The highlighted points (RSMA marked by a square and NOMA marked by a circle) represent the best fairness achieved by RSMA and NOMA respectively. 83

4.3 The Proposed Transceiver Architecture of uplink MIMO RSMA for Link-Level Simulations. 92

4.4 The MMF performance of RSMA, NOMA and SDMA versus blocklengths with three different transmit SNR averaged over 100 random channel realisations. $K = 2$. $N_t = 2$. (a) is the underloaded scenario; (b) is the overloaded scenario. 95

4.5 The MMF performance of RSMA, NOMA and SDMA in the underloaded network settings averaged over 100 random channel realisations. Blocklength is 250. $K = 4$. $N_t = 2$. (a) is the MMF versus transmit SNR; (b) is the relative gain of RSMA at the transmit SNR of 20 dB. 97

4.6 The MMF performance of RSMA, NOMA and SDMA in the overloaded network settings averaged over 100 random channel realisations. Blocklength is 250. $K = 4$. $N_t = 2$. (a) is the MMF versus transmit SNR; (b) is the relative gain of RSMA at the transmit SNR of 20 dB. 98

4.7 The max-min throughput of RSMA, NOMA and SDMA versus transmit SNR averaged over 100 random channel realisations. (a) is the underloaded scenario; (b) is the overloaded scenario. 101

5.1 The RSMA-aided MEC system with two users. 106

5.2 The successful computation probability performance of RSMA and NOMA versus the size of tasks with two different transmit SNR averaged over 100 random channel realisations. $M_2 = 5.5$k bits. (a) N=250; (b) N=500; (c) N=750; (c) N=1000. 123

5.3 The successful computation probability performance of RSMA and NOMA versus blocklength with two different task sizes averaged over 100 random channel realisations. Transmit SNR is 15 dB. 125

5.4 The successful computation probability performance of RSMA and NOMA versus transmit SNR with four different finite blocklength averaged over 100 random channel realisations. $M_1 = 7$k bits. $M_2 = 7$k bits. 126

Chapter 1

Introduction

During the past few years, the rapid progress of wireless communications has brought attention to the emergence of next-generation networks, including Fifth Generation (5G) and Sixth Generation (6G). These advancements have sparked interest from both academic and industrial communities. It is expected that 5G/6G will have the capability of meeting the growing demands for higher data rate, improved reliability, varied quality of service needs, and extensive connectivity. These advanced networks are designed to support enhanced Mobile Broadband (eMBB), Ultra-Reliable and Low-Latency Communications (URLLC), and massive Machine Type Communications (mMTC), while also adopting new approaches such as the Finite Blocklength (FBL) regime [6–8]. Mobile Edge Computing (MEC) enabling mobile devices to offload their tasks to edge servers is considered to be one promising solution for latency-critical and computation-intensive applications [9]. To address the complex requirements of future wireless networks, it is acknowledged that the advancement and refinement of Multiple Access (MA) techniques are vital since they play an important role in optimising wireless resource utilisation and enhancing interference management efficiency.

1.1 A New Multiple Access Technology: Rate-Splitting Multiple Access

Meeting the rigorous demands of URLLC and the evolving requirements of eMBB necessitates a transform from conventional MA techniques towards a transmission scheme customised to the requirements of 5G and 6G. MA techniques in wireless communication have evolved significantly over the years, with advancements occurring along both orthogonal and non-orthogonal dimensions: the orthogonal approach, which assigns dedicated resources to individual users, and the non-orthogonal approach, which enables multiple users to share the same resource [10].

Orthogonal Multiple Access (OMA) schemes, such as Time Division Multiple Access (TDMA), Frequency Division Multiple Access (FDMA), Code Division Multiple Access (CDMA) [1] and Orthogonal Frequency Division Multiple Access (OFDMA) are among the earliest techniques employed to enable multiple users to share the available spectrum efficiently for First Generation (1G), Second Generation (2G), and Third Generation (3G) wireless communication systems. TDMA divides the time resource into different time slots, allowing different users to transmit in their assigned time slots while FDMA divides the frequency spectrum into separate channels, each allocated to individual users. CDMA utilises orthogonal, user-specific codes to spread the modulated user symbols, serving multiple users simultaneously in the same time-frequency resources without causing multi-user interference (under ideal propagation conditions) [11–13]. In OFDMA data transmission is divided into multiple orthogonal subcarriers to enable simultaneous transmission from multiple users [14]. These orthogonal techniques laid the foundation for early wireless systems that aim to avoid inter-user interference and simplify transceiver design. However, the number of supported users is limited by the number of available orthogonal

[1] Here CDMA refers to CDMA with orthogonal codes. See reference [11].

resources in OMA. Consequently, it remains a challenge for these MA techniques to satisfy the Spectral Efficiency (SE)[2] and massive connectivity requirements of 5G [15, 16].

To tackle this problem, MA techniques are evolving towards improving SE by sacrificing the orthogonality of resource allocation [8]. Driven by the escalating demand for wireless communications and the limited spectrum resources available, MIMO technology has been embraced in Fourth Generation (4G) and 5G networks [17,18]. It has been studied and applied widely since it is capable of providing significantly higher SE and reliability of wireless systems by multiplexing signals in space. Unlike point-to-point MIMO where two devices with multiple antennas communicate with each other, MU-MIMO is more reliable and less sensitive to propagation environment because of MU diversity[3]. The capacity of a MIMO Gaussian Broadcast Channel (BC) can be achieved by Dirty Paper Coding (DPC) when Channel State Information at Transmitter (CSIT) is perfect which involves coding the data that align with themselves [22]. Other two non-linear precoding methods, namely Vector Perturbation (VP) [23] and lattice-aided methods [24] can also perform well in capacity.

However, despite its theoretical appeal, these non-linear precoding methods prove impractical due to the computational complexity associated with the encoding process. Besides, MIMO introduces spatial dimensions, leading to a more practical approach Space Division Multiple Access (SDMA) [25] which is proposed to maximise the capacity [26, 27] and MU-MIMO implemented using Multi-User Linear Precod-

[2]Spectral efficiency refers to the rate that can be transmitted over a given bandwidth in a specific communication system. It is a measure of how efficiently a limited frequency spectrum is utilised in bits/s/Hz. It is the bit rate divided by the bandwidth in hertz of a communication channel.

[3]Multi-User diversity, first introduced in [19] and further promoted in [20], is a form of selection diversity among users provided by independent time-varying channels across the different users. Provided that the BS is able to track the user channel fluctuations (based on feedback), it can schedule transmissions to the users with favorable channel fading conditions, i.e., near their peaks, to improve the total cell throughput [21].

ing (MU-LP). Through effective utilisation of spatial resources and multi-antenna processing at the transmitter, SDMA can concurrently serve multiple users within the same time-frequency resource. MU-LP involves employing linear precoding at the transmitter while treating multi-user interference as noise at the receivers [28]. As a result, SDMA based on MU-LP holds significant importance in numerous transmission schemes for 4G and 5G networks. In scenarios with perfect CSIT and underloaded network (i.e., when the number of antennas deployed at the transmitter is more than the number of receive antennas at all users), SDMA effectively eliminates multi-user interference if the user channels are not aligned[4]. Furthermore, SDMA achieves the maximum Degree of Freedom (DoF) of the underloaded multi-antenna BC [29]. However, SDMA is sensitive to network load, user deployment, and the accuracy of CSIT. The performance of SDMA deteriorates in overloaded scenarios (i.e., when the number of antennas deployed at the transmitter is less than the number of receive antennas at all users), requires sophisticated scheduling for optimal user pairing, and is impacted by imperfect CSIT, leading to decreased performance due to ineffective interference mitigation from imprecise linear precoding at the transmitter [30]. Practically, CSIT inaccuracies arise from various factors such as channel estimation errors, pilot contamination, feedback limitations, user mobility, and RF chain calibration errors [31].

In contrast to SDMA, where each user can directly decode the intended data streams by treating any residual interference as noise, NOMA has gained prominence in recent years, where users are superposed in the same time-frequency resources via the power domain (e.g., power-domain NOMA) [32] or code domain (e.g., sparse code multiple access) [11]. This thesis focuses on power-domain NOMA [5] which re-

[4]An aligned channel typically refers to a situation where multiple users' channels exhibit a specific spatial relationship, often characterised by high correlation or overlap in certain dimensions. SDMA relies on the spatial diversity of user channels and utilises beamforming and advanced signal processing to manage interference. Aligned channels would reduce the effectiveness of SDMA by making it harder to separate users spatially.

[5]In this thesis, power-domain NOMA will be referred to NOMA.

lies on Superposition Coding (SC) [6] at the transmitter to superpose user messages in the power domain and SIC at the receivers (denoted in short as Superposition Coding-Successive Interference Cancellation (SC-SIC)) [33,34]. A significant benefit of NOMA lies in its ability to enhance SE in heavily overloaded scenarios by serving users with closely aligned channels and varying channel strengths within the same time-frequency resources. However, NOMA, originally designed for single-antenna setups, leads to suboptimal utilisation of spatial dimensions, resulting in a loss of multiplexing gains and data rate in multi-antenna configurations [35]. The drawbacks of NOMA are outlined as follows. First, the deployment of SC-SIC in NOMA originates from the concept of a degraded BC where users are ordered by their channel strengths. However, the multi-antenna BC is non-degraded, and users cannot be ordered based on channel strengths. Applying SC-SIC to a non-degraded BC results in a loss in DoF [35]. Second, the effectiveness of NOMA is limited to specific scenarios where users are closely aligned with a disparity of channel strengths. Third, implementing NOMA in multi-antenna setups increases complexity at both the transmitter and receivers, i.e. joint optimisation of precoding, user grouping, and decoding sequences due to the need for multi-layer SIC and the lack of a natural user channel order. Fourth, multi-antenna NOMA is sensitive to imperfect CSIT.

These non-orthogonal techniques offer enhanced SE, improved system capacity, and better support for diverse traffic types, making them well-suited for next-generation wireless networks. Nevertheless, SDMA and NOMA can be considered as two extreme interference management strategies, i.e., SDMA relies on fully treating any residual multi-user interference as noise, while NOMA relies on some users to fully decode and cancel interference from other users.

Conversely, RSMA emerges as a novel solution that circumvents the aforementioned limitations. RSMA utilises linear precoding with Rate-Splitting (RS) at the trans-

[6]Superposition coding refers to superposition coding which is a technique to combine different streams by adding them together with different power levels.

mitter and SIC at the receivers. RSMA involves splitting the message of a user into common and private parts, enabling each part to be decoded by one or more receivers. To reconstruct the original message, a receiver must retrieve both the common part and the corresponding private part [36, 37]. A significant advantage of RSMA is its ability to flexibly manage inter-user interference. This allows for the decoding of partial interference while treating the remaining as noise [31]. Such capability enables RSMA to bridge the gap between these two extreme interference management strategies of fully decoding interference and treating interference as noise [38]. Depending on the particular methods employed for message splitting and combining, RSMA encompasses a variety of transmission schemes. These include linearly precoded 1-layer RS, 2-layer Hierarchical RS (HRS), and Generalised RS (GRS), each of which can be further expanded to their respective nonlinear counterparts [8, 38, 39], and even to space-time designs [40]. Through analytical investigations in information and communication theory, RSMA has been demonstrated to be optimal in terms of the DoF for various transmission scenarios [41, 42]. Compared to traditional MA techniques, RSMA offers enhancements in Energy Efficiency (EE)[7] [43–45], user fairness [46–48], and Quality-of-Service (QoS) [38] across a broad spectrum of network loads, encompassing both underloaded and overloaded scenarios, as well as diverse user channel conditions. Additionally, RSMA demonstrates improved robustness to inaccuracies in CSIT [49], while also requiring lower feedback overhead and complexity [50].

All aforementioned works focus on downlink RSMA in multi-antenna multiple-access scenarios. However, it has been demonstrated that uplink RSMA can achieve the optimal rate region of an K-user Gaussian Multiple Access Channel (MAC) using up to $2K - 1$ virtual point-to-point Gaussian channels created by message splitting [51, 52]. Comparative analyses with traditional uplink communication

[7]Energy efficiency is typically defined as the ratio of the data rate to the total power consumed by the communication system. The unit is bits per joule (bits/J).

schemes, namely NOMA and OMA, have revealed that uplink RSMA in Single-Input Single-Output (SISO) channels yields superior performance in terms of sum-rate, sum-throughput [53–55], outage performance [56]. Besides, uplink RSMA holds the capability to achieve the capacity region without the need of time-sharing among users which requires stringent synchronisation due to the coordination of the transmissions of all users [57, 58]. In [59], uplink RSMA demonstrates enhanced user fairness compared to NOMA in Single-Input Multiple-Output (SIMO) channel. Furthermore, the investigation of uplink MIMO RSMA under electromagnetic exposure constraints [60] highlights significant performance advantages over NOMA and SDMA.

Overall, RSMA has emerged as a general, universal, flexible, robust and powerful framework for the design and optimisation of non-orthogonal transmission, MA, and interference management strategies. Driven by the appealing advantages of uplink RSMA in wireless communications, the main objective of this thesis is to investigate the performance of uplink RSMA in the FBL regime, which can contribute to the evolution of next-generation wireless networks.

1.2 Motivation

URLLC stands as a cornerstone of 5G and future 6G networks, aiming to meet the stringent requirements for highly reliable and low-latency communication[8]. These demands are pivotal for supporting critical applications in fields such as MEC. In this context, systems are compelled to operate with short packets to meet the demands of URLLC which necessitates the adoption of FBL codes techniques to minimise latency while maintaining high reliability.

Nevertheless, the Shannon capacity theory states that there exists a maximum rate at which information can be transmitted reliably with arbitrarily low decoding error probability and infinite blocklength, so it becomes inadequate for URLLC systems, primarily owing to FBL transmission characteristics. With the consideration of FBL, the study in [61] has paved the way for theoretically analysing the performance with FBL codes by laying out the fundamental bounds for the achievable rate with a given blocklength and error probability in Additive White Gaussian Noise (AWGN) channels. This research establishes the foundation for examining and developing short-packet communications. To achieve reduced latency for the implementation of URLLC in 5G and 6G, it is crucial to investigate the performance of emerging physical-layer MA techniques in the FBL regime.

In this thesis, the emphasis is placed on the MA scheme, namely, RSMA, especially in the uplink channel. It has garnered significant interest in recent years due to its ability to improve system performance across various network loads and user channel conditions [8, 31, 38].

[8]Latency refers to the time delay between the initiation of a transmission and the successful reception of the data. This encompasses several components such as processing delay, queueing delay, transmission delay, propagation delay and retransmission delay. In the context of URLLC, the latency is required to be extremely low, often less than 10 ms.

1.3 Organisation and Contributions

As all research working on uplink RSMA is based on the assumption of Infinite Blocklength (IFBL) [51, 52, 54, 56, 59, 62, 63], the primary aim of this thesis is to explore the performance of uplink RSMA in the FBL regime since URLLC is acknowledged as a critical factor in shaping the landscape of next-generation wireless networks.

In this section, summaries of each chapter, outlining the main contributions, are provided below.

- **Chapter 2 Preliminaries:** A background overview of the fundamental and related literature of this thesis are provided by presenting the development of the conventional multiple access technologies, followed by the principles RSMA from both downlink and uplink aspects, and state-of-the-art research on FBL and MEC.

- **Chapter 3 SISO RSMA in Finite Blocklength Regime:** For the *first time*, the concept of uplink RSMA is introduced into short-packet communications. This work differs from previous works on uplink RSMA by addressing practical scenarios of data transmission with FBL code, which are essential for low-latency communications in 5G, URLLC and future networks.

  The performances of sum-throughput regarding error probability, and rate region are investigated with the consideration of a two-user uplink SISO channel system. The optimal power allocation and rate allocation factors that maximise the sum-throughout with FBL constraints are designed. The impacts of blocklength and target transmission rate on the throughput and error probability performance of uplink RSMA where one user message is split are investigated. Numerical results demonstrate the superiority of RSMA, i.e., RSMA

can achieve higher throughput with smaller blocklength than conventional MA techniques, resulting in lower latency and higher reliability.

- **Chapter 4 MIMO RSMA in Finite Blocklength Regime:** For the *first time*, an uplink MIMO RSMA framework is proposed and both precoders and combiners with Max-Min Fairness (MMF) metric[9] under FBL constraints are optimised. Alternating Optimisation (AO) is employed to decompose the problem into two subproblems, given the strong coupling between precoders and combiners. Additionally, a low-complexity scheme is introduced to design the decoding order at the receiver. Subsequently, the Physical (PHY)-layer of the uplink MIMO RSMA architecture is designed. The MMF gains over conventional transmission schemes such as SDMA and NOMA and the robustness to network loads are showcased by the proposed uplink MIMO RSMA schemes through numerical evaluations. Link Level Simulations (LLS) results confirm the improved max-min throughput benefits of RSMA over SDMA and NOMA.

- **Chapter 5 RSMA with Mobile Edge Computing:** In this chapter, the RSMA scheme is proposed to assist an MEC system in FBL regime where local computation tasks from two users are partially offloaded and processed at the MEC server. An RSMA-aided MEC framework is proposed and the overall Successful Computation Probability (SCP) [10] of the service is derived. The offloading factor, power allocation and task-splitting factor are optimised with the metric of MEC under FBL constraints. AO is applied to decompose the optimisation problem into subproblems which are solved by Successive Convex Approximation (SCA)-based approaches. Numerical results demonstrate that

[9]MMF is a resource allocation strategy that aims to maximise the worst performance (e.g., data rate, throughput) across all users in the network.

[10]SCP is defined as the probability of the case that all the offloaded tasks and the remaining local tasks are computed successfully at the MEC server and the users respectively within the delay budget. Thus SCP contains two parts which are the successful offloading probability and the successful execution probability, respectively.

applying uplink RSMA in the MEC system with FBL constraints can not only improve the SCP performance but also provide lower latency in comparison to NOMA.

- **Chapter 6 Conclusions:** In this chapter, the achievements of this thesis and explore potential directions for future research are summarised.

1.4 Publications

The work in this thesis has resulted in a number of papers, that have been published.

Published Journals

- **J. Xu**, O. Dizdar and B. Clerckx, "Rate-Splitting Multiple Access for Short-Packet Uplink Communications: A Finite Blocklength Analysis," in IEEE Communications Letters, vol. 27, no. 2, pp. 517-521, Feb. 2023, doi: 10.1109/LCOMM.2022.3226817. [11]

- **J. Xu** and Y. Mao, "Max-Min Fairness for Uplink Rate-Splitting Multiple Access with Finite Blocklength," in IEEE Transactions on Vehicular Technology, doi: 10.1109/TVT.2025.3528938. [12]

- **J. Xu**, and B. Clerckx, "Max-Min Fairness and PHY-Layer Design of Uplink MIMO Rate-Splitting Multiple Access with Finite Blocklength," in IEEE Transactions on Communications, doi: 10.1109/TCOMM.2024.3490492. [13]

[11] This paper is related to the Chapter 3.
[12] This paper introduces the foundational concepts explored in Chapter 4.
[13] This paper is related to the Chapter 4.

Chapter 2

Preliminaries

This chapter presents the background knowledge and state-of-the-art works covered in this thesis, including the fundamentals of RSMA from both downlink and uplink channel aspects, FBL regime and MEC system.

2.1 Fundamentals of RSMA

In this section, the fundamentals of RSMA in both downlink and uplink channels are introduced, considering a K-single-antenna-user, indexed by $\mathcal{K} = \{1, ..., K\}$, served by a BS equipped with N_t transmit antennas system.

2.1.1 Downlink RSMA

For downlink 1-layer RSMA, which is the simplest and most practical RSMA implementation shown in Fig. 2.1 [8], the message intended for user-k is represented by W_k, $\forall k \in \mathcal{K}$. Each W_k is split into two parts, denoted as $\{W_{c,k}, W_{p,k}\}$ and known as the common part and the private part, respectively. The common parts are combined into a common message $W_c = \{W_{c,1}, \cdots, W_{c,K}\}$, which is then encoded into

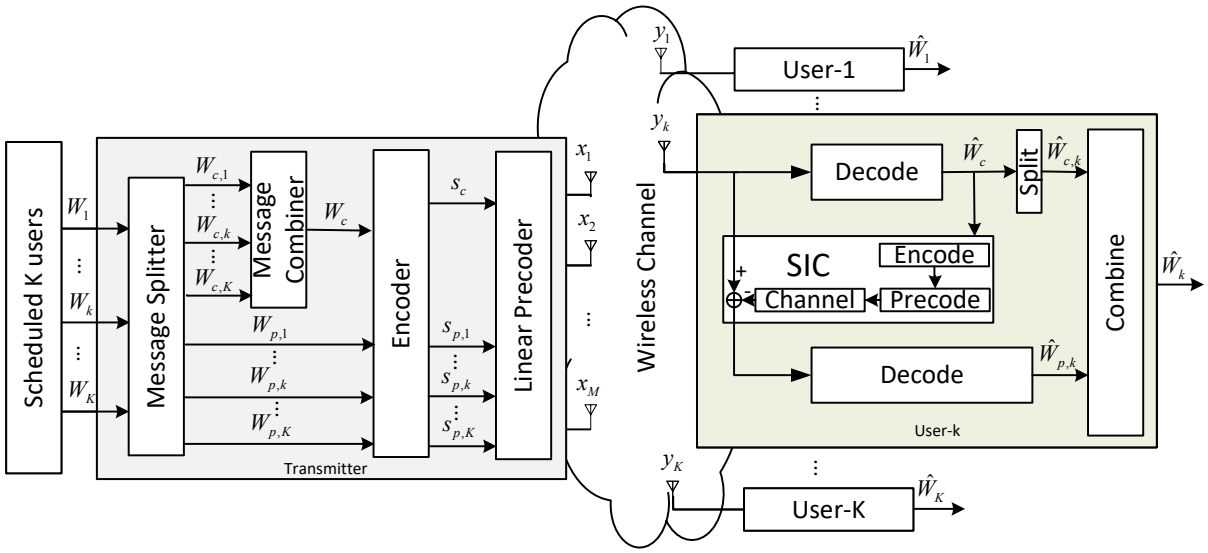


Fig. 2.1: Transceiver architecture of K-user downlink RSMA. Message splitter is implemented by demultiplexer.

a single common stream s_c. The private messages are encoded into corresponding private streams $s_k,\ \forall k \in \mathcal{K}$.

The linear precoders/beamforming vectors for common and private streams are denoted as $\mathbf{p}_c \in \mathbb{C}^{N_t \times 1}$ and $\mathbf{p}_k \in \mathbb{C}^{N_t \times 1},\ \forall k \in \mathcal{K}$, respectively. The transmit signal $\mathbf{x} \in \mathbb{C}^{N_t \times 1}$ can be expressed as

$$\mathbf{x} = \mathbf{P}\mathbf{s} = \mathbf{p}_c s_c + \sum_{k=1}^{K} \mathbf{p}_k s_k, \tag{2.1}$$

where $\mathbf{P} = [\mathbf{p}_c, \mathbf{p}_1, \cdots, \mathbf{p}_K]$ is the precoding matrix. The vector of transmitted symbol is $\mathbf{s} = [s_c, s_1, \cdots, s_K]^T \in \mathbb{C}^{(K+1)\times 1}$. Assume $\mathbb{E}\{\mathbf{s}\mathbf{s}^H\} = \mathbf{I}_{K+1}$, then the sum transmit power constraint is $\mathrm{tr}\left(\mathbf{P}\mathbf{P}^H\right) \leq P$, where P is the transmit power budget. The received signal at user-k is given by

$$\begin{aligned} y_k &= \mathbf{h}_k^H \mathbf{x} + n_k \\ &= \mathbf{h}_k^H \mathbf{p}_c s_c + \mathbf{h}_k^H \sum_{k=1}^{K} \mathbf{p}_k s_k + n_k, \end{aligned} \tag{2.2}$$

where $\mathbf{h}_k \in \mathbb{C}^{N_t \times 1}$ denotes the channel vector between the k-th user and the BS. $n_k \sim \mathcal{CN}\left(0, \sigma_{n,k}^2\right)$ is AWGN of zero mean and variance $\sigma_{n,k}^2$ at the receiver. It is assumed that the variance at each receiver is the same, i.e. $\sigma_{n,1}^2, \ldots, \sigma_{n,K}^2 = \sigma_n^{21}$.

At the receiver side, each user first decodes the common stream s_c by treating the interference from the private streams $s_1, \ldots, s_K$ as noise. Accordingly, the Signal-to-Interference-plus-Noise-Ratio (SINR) of decoding s_c at user-k is expressed as

$$\rho_{c,k} = \frac{|\mathbf{h}_k^H \mathbf{p}_c|^2}{\sum_{j=1}^{K} |\mathbf{h}_k^H \mathbf{p}_j|^2 + \sigma_n^2}. \tag{2.3}$$

With the assumption of successfully decoding the common stream, s_c is reconstructed and subtracted from the received signal y_k. The private stream s_k is decoded subsequently at user-k with the private streams of other users treated as noise. Then, the SINR of the decoded private stream s_k at user-k is expressed as

$$\rho_{p,k} = \frac{|\mathbf{h}_k^H \mathbf{p}_k|^2}{\sum_{j \in \mathcal{K}, j \neq k} |\mathbf{h}_k^H \mathbf{p}_j|^2 + \sigma_n^2}. \tag{2.4}$$

After the common and private streams are decoded, the original message intended for user-k is reconstructed by extracting the decoded $\widehat{W}_{c,k}$ from the decoded $\widehat{W}_c$, and combining the decoded $\widehat{W}_{c,k}$ with the decoded $\widehat{W}_{p,k}$. Assuming Gaussian signaling and infinite blocklength, the achievable rates for decoding the common and private streams at user-k are

$$R_{c,k} = \log_2\left(1 + \rho_{c,k}\right), \tag{2.5a}$$

$$R_{p,k} = \log_2\left(1 + \rho_{p,k}\right). \tag{2.5b}$$

The unit of Eq. (2.5a) and Eq. (2.5b) is bit/s/Hz. To ensure the common stream s_c

[1]The noise at each user is assumed to be Additive White Gaussian noise of zero mean and variance σ_n^2. That is, for the k-th user, the variance is $\sigma_n^2 = \sigma_{n,k}^2$ for every k.

is successfully decoded by all users, its rate cannot exceed $R_c = \min_{\forall k} R_{c,k}$. Since s_c contains sub-messages $W_{c,1}, \ldots, W_{c,K}$ of all the K users, let C_k denote the portion of rate R_c allocated to user-k for $W_{c,K}$. Then, the constraint about the common rate is written as $R_c = \sum_{k \in \mathcal{K}} C_k$. As a consequence, the overall achievable rate of user-k is written as

$$R_{k,tot} = C_k + R_{p,k}. \tag{2.6}$$

The precoding matrix $\mathbf{P}$ can be designed using low-complexity techniques, such as zero-forcing beamforming (ZFBF) for the private streams and random beamforming for the common stream [46], or through optimisation strategies aimed at various objectives, i.e. Weighted Minimum-Mean Square Error (WMMSE), including maximising the Weighted Sum Rate (WSR) [38, 42], maximising the minimum user rate [64], enhancing EE [43], minimising transmit power [65], etc.

In this section, the fundamentals of downlink 1-layer RSMA is introduced and interested readers are encouraged to refer to [38] for a more comprehensive study on alternative forms of RSMA. 1-layer RSMA necessitates only one layer of SIC at each receiver. Unlike the GRS described in [38], in which multiple common streams and multiple SIC layers are involved at the receivers, 1-layer RSMA avoids the need for user grouping and ordering. Consequently, encoding complexity, scheduling complexity and receiver complexity are significantly reduced. Results in [38] demonstrate that the low-complexity 1-layer RSMA achieves comparable rate performance to the GRS. This advantage in complexity reduction becomes particularly pronounced as the number of users increases.

In the literature, downlink RSMA has been proven to be a strong multi-user transmission scheme and reliable interference management strategy for multi-antenna wireless networks. It has the potential to tackle numerous challenges of modern communication systems. Downlink RSMA has been shown to bridge and outperform

existing multiple access schemes, such as SDMA, NOMA, OMA and multicasting under perfect [36,38,66] and imperfect CSIT [42,64,67–70]. In the downlink channel, the benefit of RSMA lies in its flexibility to partially decode the multi-user interference and partially treat it as noise; in contrast with SDMA, which fully treats the interference as noise, and NOMA, which fully decodes the interference [31,36,38,67]. More specifically, based on this flexibility, [38] has demonstrated that RSMA exhibits greater robustness to influential factors such as channel disparity, channel orthogonality, network load, and the quality of CSIT. [42] and [64] have explored the sum-DoF and Max-Min Fairness-DoF of underloaded MU-Multiple-Input Single-Output (MISO) systems under imperfect CSIT. Moreover, [71] has established that RSMA achieves superior sum-rate performance compared to SDMA and NOMA in imperfect CSIT and Channel State Information at Receiver (CSIR) scenarios. Notably, the benefits of RSMA extend beyond Gaussian signaling and infinite block-length assumptions, as evidenced by practical system implementations [72] and throughput enhancements observed through link-level simulations [73–75].

Due to its intrinsic message splitting capability, which distinguishes it from other MA schemes, the advantages of downlink RSMA can be summarised below:

- partially decoding interference and partially treating interference as noise,
- subsuming SDMA, NOMA, OMA, and multicasting as sub-schemes.

2.1.2 Uplink RSMA

RSMA has been studied not only for the downlink channel but also for uplink channel [51,52,54,56,59,62,63]. Uplink RSMA was first proposed in [51,52] for a K-user Gaussian MAC and it has proven that the capacity region of a Gaussian MAC can be achieved by creating $2K-1$ virtual channels through message splitting with-

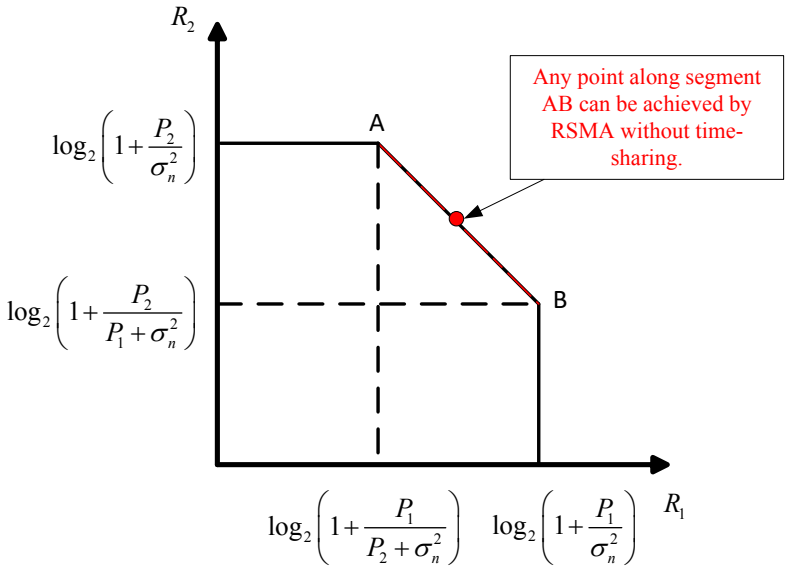


Fig. 2.2: Capacity region of the two-user uplink AWGN channel. The two corner points, A and B, correspond to the maximum rates achieved by two SIC decoding orders: A is achieved by decoding user-1 first and B is achieved by decoding user-2 first. P_1 and P_2 is the transmit power for user-1 and user-2, respectively. σ_n^2 is the noise variance at the receiver.

out the need of time-sharing[2]. The capacity region of the two-user SISO Gaussian MAC is a pentagon as shown in Fig. 2.2. The two corner points A and B can be achieved by SIC with two reverse decoding orders. Nevertheless, the rate points along the segment AB cannot be achieved with the same method. There exist three methods capable of achieving the points on the segment AB, which are joint encoding/decoding, time-sharing, and rate splitting. However, joint encoding/decoding is impractical due to the high complexity involved in code construction and joint decoding at the receiver [51, 76]. The second method, time-sharing, requires stringent synchronisation due to the need to coordinate the transmissions of all users [51].

Compared to the aforementioned two methods, uplink RSMA turns out to be a more attractive approach since every point of the capacity region including the segment AB can be achieved via SC-SIC. The transceiver architecture of the K-

[2]Time-sharing is a resource allocation and management method widely used in wireless communication. The concept involves dividing time into multiple slots, where different users occupy the resource exclusively during their allocated time slot, ensuring efficient resource sharing.

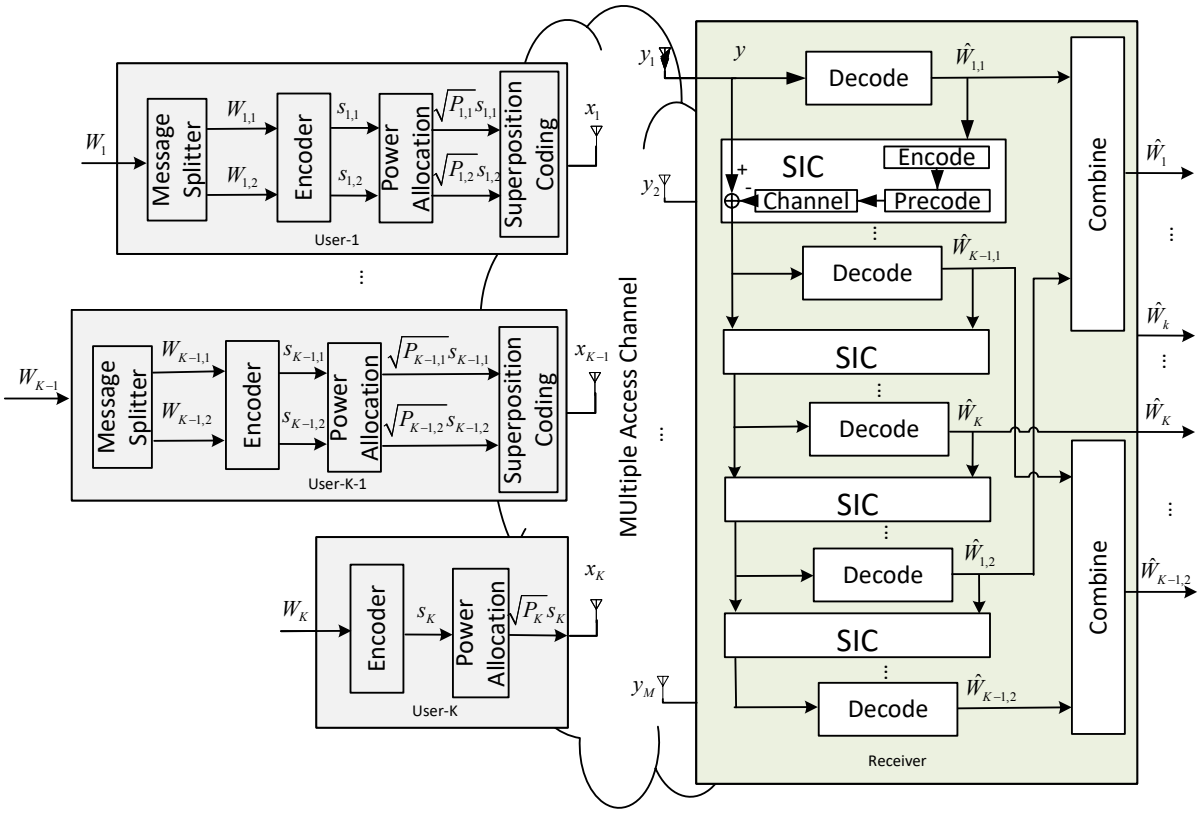


Fig. 2.3: Transceiver architecture of K-user uplink RSMA.

user uplink RSMA system is illustrated in Fig. 2.3. Without loss of generality, it is assumed that the messages of all users except user-K are split. Then, SIC technique is applied at the BS to decode the messages of all users [51]. The message of user-k, $k \in \{1, ..., K-1\}$ is split into two parts as $W_k = \{W_{k,1}, W_{k,2}\}$ and then they are independently encoded into $\mathbf{s}_k = [s_{k,1}, s_{k,2}]$, where $\mathbb{E}\{|s_{k,i}|^2\} = 1$, $i = \{1, 2\}$. The two streams are allocated with power $P_{k,1}$ and $P_{k,2}$, respectively, and superposed at user-k. The message W_K of user-K is directly encoded into s_K and allocated with power P_K. The transmitted signal at user-k is written as

$$x_k = \begin{cases} \sqrt{P_{k,1}} s_{k,1} + \sqrt{P_{k,2}} s_{k,2}, & k \in \{1, ..., K-1\} \\ \sqrt{P_k} s_k, & k = K. \end{cases} \tag{2.7}$$

Then the received signal at the BS is given by

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{h}_k x_k + \mathbf{n}, \tag{2.8}$$

where $\mathbf{h}_k \in \mathbb{C}^{N_t \times 1}$ is the channel vector between the user-k and the BS, and $n \sim \mathcal{CN}\left(\mathbf{0}_{N_t}, \sigma_n^2 \mathbf{I}_{N_t}\right)$ is AWGN noise at the BS. Without loss of generality, it is assumed that the noise variance is one, i.e., $\sigma_n^2 = 1$. At the BS, $\mathbf{g}_{k,i} \in \mathbb{C}^{N_t \times 1},\ i \in \{1, 2\}$ are used to detect two streams from user-k, $k \in \{1, ..., K-1\}$ while $\mathbf{g}_K \in \mathbb{C}^{N_t \times 1}$ is applied to detect the stream from user-K. Let η denote the decoding order of the $2K-1$ received streams $\{s_{k,i}, s_K \mid k \in \{1, ..., K-1\},\ i \in \{1, 2\}\}$ and $\eta_{k,i}$ denotes the decoding order of $\eta_{k,i}$ with stream $\eta_{k,i}$ being decoded before stream $s_{k',i'}$ if $\eta_{k,i} < \eta_{k',i'}$. For user-$K$, $\eta_{K,i}$ is simplified to η_K. With the assumption of Gaussian signalling and infinite blocklength, the rate expressions of decoded streams are given below

$$R_{k,i} = \log_2 \left(1 + \frac{P_{k,i} |\mathbf{g}_{k,i}^H \mathbf{h}_k|^2}{\sum_{\eta_{k',i'} > \eta_{k,i}} P_{k',i'} |\mathbf{g}_{k,i}^H \mathbf{h}_k|^2 + \sigma_n^2} \right) \tag{2.9a}$$

$$R_K = \log_2 \left(1 + \frac{P_K |\mathbf{g}_K^H \mathbf{h}_k|^2}{\sum_{\eta_{k',i'} > \eta_K} P_{k',i'} |\mathbf{g}_K^H \mathbf{h}_k|^2 + \sigma_n^2} \right). \tag{2.9b}$$

The sum-rate maximisation by optimising the power allocation $\{P_{k,i}, P_K \mid k \in \{1, ..., K-1\},\ i \in \{1, 2\}\}$ has been investigated in [54]. To decode all $2K$-1 streams, the receiver necessitates $2K$-2 layers of SIC. Consequently, opting to avoid time sharing introduces increased receiver complexity. The system model described above for uplink RSMA is a general case and can be customised to facilitate uplink multiple access for both homogeneous and heterogeneous services.

It has been demonstrated that uplink RSMA can not only improve the capacity in SISO channel [54] but also the MMF in SIMO channel [59]. In [59], considering an MMF problem in an uplink SIMO setting, SIMO-RSMA has been shown

to achieve a higher minimum data rate and lower latency than SIMO-OMA and SIMO-NOMA. The metric of MMF has also been investigated in a cooperative-rate-splitting-multiple-access (C-RSMA) scheme in [77], showing C-RSMA has a better performance rather than C-NOMA, and C-NOMA has a better performance compared with C-OMA. The closed-form of outage probability is derived in [55] based on slotted ALOHA and RSMA. Moreover, uplink RSMA can be employed to enable a specific service in a network, such as URLLC or mMTC, i.e. uplink RSMA with network slicing has been stated to provide a more flexible decoding order and a larger achievable rate region than OMA and NOMA [57]. The cooperation of uplink RSMA and multibeam satellite has been proposed to improve the SE with imperfect Channel State Information (CSI) [78]. A Reconfigurable Intelligent Surface (RIS)-assisted uplink two-user RSMA system has been investigated in terms of the maximum achievable rate of one user subject to another user's QoS requirements [79]. Uplink RSMA with multiple antennas has also been studied in [60] that RSMA with electromagnetic exposure constraints demonstrate the significant gain over NOMA and SDMA in terms of sum-rate.

Although uplink RSMA has been studied, the research in this area is relatively limited and all existing studies are based on the essential assumption of IFBL. Therefore, further investigation into uplink RSMA is still warranted, particularly when considering practical FBL conditions. Another interesting topic is to study for uplink RSMA, how many users' messages should be split to have a relatively high gain compared to conventional transmission schemes as the number of splitting users can cause an increased complexity at the receiver side.

2.2 Finite Blocklength Regime

The remarkable establishment of an abstract communication model proposed by Shannon in [80] transforms numerous practical engineering problems into precisely formulated mathematical challenges. With the consideration of a constant noise level and error-free transmission, a theoretical upper bound on the maximum achievable data rate (capacity) of a communication channel is calculated by the Shannon capacity formula. This capacity describes that any rate inside the capacity region can be achieved with arbitrarily small error probability under the assumption of IFBL.

However, it is agreed that the 5G communication technology needs to be capable of supporting three generic services: mMTC, URLLC, and eMBB. Some typical use cases are shown in Table 2.1 [1, 81–83]. Among these three services, URLLC is required to provide the service with the highest reliability, i.e. 99.999%, and the lowest latency, i.e. less than 1 ms [1, 2]. The importance of key capabilities in different usage scenarios is shown in Table 2.2 [3].

In this sense, systems must function with short-packet due to the low-latency requirements of URLLC, which makes it necessary to analyse system performance

Applications \ Requirements		User experienced data rates	Latency	Reliability
eMBB	Uplink file transfer	50 Mbps	16 s	99.9%
	AR/VR	50 Mbps	10 ms	99 %
URLLC	Factory automation	10 Mbps	10 ms	99.99%
	Motion control	1 Mbps	1 ms	99.9999%
mMTC	Sensing and factory logistics	1 Mbps	0.5 s	95%
	Wearables	2-5 Mbps	Relaxed	N/A

Table 2.1: Requirements of eMBB, URLLC and mMTC applications [1, 2]

The importance of key capabilities in different usage scenarios			
	eMBB	**mMTC**	**URLLC**
Peak data rate	⋆⋆⋆	⋆	⋆
User experienced data rate	⋆⋆⋆	⋆	⋆
SE	⋆⋆⋆	⋆	⋆
Mobility	⋆⋆⋆	⋆	⋆⋆⋆
Latency	⋆⋆	⋆	⋆⋆⋆
Connection density	⋆⋆	⋆⋆⋆	⋆
EE	⋆⋆⋆	⋆⋆	⋆
Area traffic capacity	⋆⋆⋆	⋆	⋆

Table 2.2: The importance of key capabilities in different usage scenarios [3]

using FBL codes. URLLC demands FBL techniques to reduce latency while ensuring high reliability. To support this major application which requires low-latency for uplink MA in 5G and Beyond Fifth-Generation (B5G), short-packet communication with FBL codes is applied to reduce the transmission latency.

Compare to the Shannon's capacity based on IFBL assumption, Block Error Rate (BLER) cannot be neglected in the FBL regime [7]. In the pioneering work [61], the authors have laid the fundamental-theoretic limits for the achievable rate with FBL codes for a given blocklength and an error probability, leading the path for theoretical performance analysis with FBL codes. The achievable rate of a single-user in AWGN fading channel system is given by

$$R_{AWGN} = \log_2(1+\gamma) - \sqrt{\frac{V_{AWGN}(\gamma)}{N}} Q^{-1}(\epsilon) + \mathcal{O}\left(\frac{\log N}{N}\right) \tag{2.10a}$$

$$\approx \log_2(1+\gamma) - \sqrt{\frac{V_{AWGN}(\gamma)}{N}} Q^{-1}(\epsilon), \tag{2.10b}$$

where γ is the Signal-to-Noise Ratio (SNR), N is the blocklength, and Q is the Q-function [3]. $\mathcal{O}\left(\frac{\log N}{N}\right)$ represents higher-order terms that become less significant as N increases. The coefficient V_{AWGN} is referred to as the channel dispersion which

[3]Q-function is the tail distribution function of the standard normal distribution. Normally, Q-function is defined as: $\mathrm{Q(x)}=\frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-\frac{u^2}{2}}du$.

is written as

$$V_{AWGN}(\gamma) = \left(1 - \frac{1}{(1+\gamma)^2}\right) \log_2^2(e). \tag{2.11}$$

The channel capacity is the expectation of information density, while the channel dispersion measures the stochastic variability of the channel relative to a deterministic channel with the same capacity. The approximation (2.10) suggests that to maintain the desired error probability ϵ at the given blocklength N, there is a rate penalty proportional to $1/\sqrt{N}$ compared to the channel capacity with IFBL.

Motivated by [61], the impact of FBL on the performance of various point-to-point communication scenarios has been investigated in [84–87]. A new general upper and lower bounds to the minimum rate at a given blocklength considering Shannon's lossy source coding theorem has been given in [84]. The investigations in [86] have delved into the maximum channel coding rate within quasi-static fading channels and AWGN channels, with the constraints of BLER, blocklength, and long-term transmit power. [88] has derived the overall error probability and investigated the capacity of a single relay system under FBL conditions. Furthermore, the maximum achievable transmission rate over quasi-static MIMO fading channels has been examined in [87], exploring scenarios considering both perfect and imperfect CSIT/CSIR. The throughput achieved in cognitive radio channels with FBL codes under buffer limitations has been studied in [85]. The closed-form of average throughput in cooperative Internet of Things (IoT) networks with short packet communication has been derived in [89].

Additionally, Age of Information (AoI) is a crucial metric to measure the time elapsed since the last successfully received update packet was generated and quantify the freshness of the information being communicated [90] used in systems where timely delivery of information is essential, such as in real-time monitoring, autonomous driving, and industrial automation. Due to the character of latency re-

duction, the AoI performance of short-packet communications has been investigated to evaluate the influence of short packets on the freshness of transmitted information [91–93]. In particular, [91] has delved into the effect of packet blocklength on delay and peak AoI in a single-user system. Based on the same system, [92] has explored the probability of peak-age violation surpassing a set threshold. Furthermore, investigations into Decode-and-Forward (DF) relaying systems have examined the impacts of packet generation rates, blocklength, and allocation factors on average AoI [93].

However, the approximation (2.10) for AWGN channels holds only in case the interference is zero, i.e., there is a single user or when the transmitter has perfect CSI [94]. Under imperfect CSI, each receiver experiences interference from the signals intended for other users. To capture the impact of interference within FBL coding, [94] examined a scenario involving K sender-receiver pairs transmitting simultaneously, employing independent and identically distributed (i.i.d) Gaussian codebooks. A second-order expression for the achievable coding rate is derived as follows:

$$R_{iid} \approx \log_2(1+\gamma) - \sqrt{\frac{V_{iid}(\gamma)}{N}} Q^{-1}(\epsilon), \tag{2.12}$$

with channel dispersion

$$V_{iid}(\gamma) = \frac{2\gamma}{1+\gamma} \log_2^2(e). \tag{2.13}$$

URLLC mandates strict latency constraints on the physical layer. Besides, when multiple URLLC users connect to the same BS, resource allocation becomes crucial due to resource constraints. Therefore, studying MA techniques with FBL constraints becomes imperative in this context. Considering the advantages of NOMA over OMA in terms of the capacity region under IFBL assumption [95], the investigation of NOMA performance in the FBL regime has become an interesting topic [58, 96–105].

The trade-off between transmission rate and error probability has been investigated in [96], demonstrating that optimising the individual user's error probability results in high throughput. Motivated by this trade-off, NOMA was introduced into short-packet downlink communications and the error probabilities have been derived for NOMA in AWGN channel and Rayleigh fading channel with FBL constraints in [97, 98, 104, 105]. They have shown the superiority of NOMA over OMA in terms of effective throughput, latency, fairness and SE, respectively. In [99], a resource allocation problem about subcarrier assignment, transmission power allocation, and transmission rate adaptation for NOMA in the FBL regime has been formulated to maximise total weighted effective throughput. [100] has addressed the EE problem in downlink channel transmission systems with latency constraints. [100] has proposed a hybrid transmission scheme aimed at minimising energy consumption, combining TDMA and NOMA. [101] has investigated subchannel allocation and power control policies to maximise achievable effective EE in uplink NOMA-based mMTC networks, considering the FBL constraints. [58, 102, 103] have studied the queueing delay of NOMA compared to OMA in the uplink channel with FBL and derived the closed-form approximations for the error probability.

Although extensive research has examined the superiority of NOMA over traditional MA approaches in the context of FBL, it has been shown that NOMA is not an efficient strategy in multi-user multi-antenna systems when considering multiplexing gains and the use of SIC receivers under the assumption of IFBL [35]. On the other hand, downlink RSMA has emerged as a promising multiple access technique for multi-antenna networks in scenarios with IFBL [31] as it bridges and outperforms SDMA, NOMA, and PHY-layer multicasting in multi-antenna networks across various network loads and user deployments [36, 38]. Moreover, the work of downlink RSMA in the FBL regime has been studied in [106, 107], showing that the performance of sum-rate and MMF has been notably improved by RSMA compared to SDMA in both underloaded and overloaded cases.

Nevertheless, all existing works on RSMA with the constraint of FBL are based on the downlink channel, leaving the investigation of uplink RSMA in the FBL regime as a research problem.

2.3 Mobile Edge Computing

Over the past decade, the widespread use of mobile devices and the exponential surge in mobile Internet traffic[4] have driven significant advancements in wireless communications and networking. Notably, innovations in small-cell networks, multi-antenna systems, and millimetre-wave technologies can potentially deliver gigabit wireless access in next-generation networks [108]. These high-rate and highly reliable air interfaces enable mobile devices to offload computing tasks to remote cloud data centres, giving rise to the field of Mobile Cloud Computing (MCC). However, MCC faces a critical limitation—the long propagation distance between end users and remote cloud centres, which introduces substantial latency for mobile applications. It is widely acknowledged that relying only on Cloud Computing is insufficient to achieve the millisecond-level latency targets required for computing and communication in 5G networks. Additionally, the massive data exchange between end users and remote cloud servers risks overwhelming and congesting networks, potentially leading to network bottlenecks. To address this challenge, it is crucial to complement Cloud Computing with MEC, which offloads traffic, computation, and network functions closer to the network edge.

The concept of MEC was firstly proposed by the European Telecommunications Standard Institute (ETSI) in 2014, and was defined as a new platform that "provides IT and cloud computing capabilities within the Radio Access Network (RAN) in close proximity to mobile subscribers" [109]. Recently, there is an increasing number

[4]The increasing volume of data being transmitted over wireless networks due to the demand for high-quality services, high data rates, and advanced user terminals.

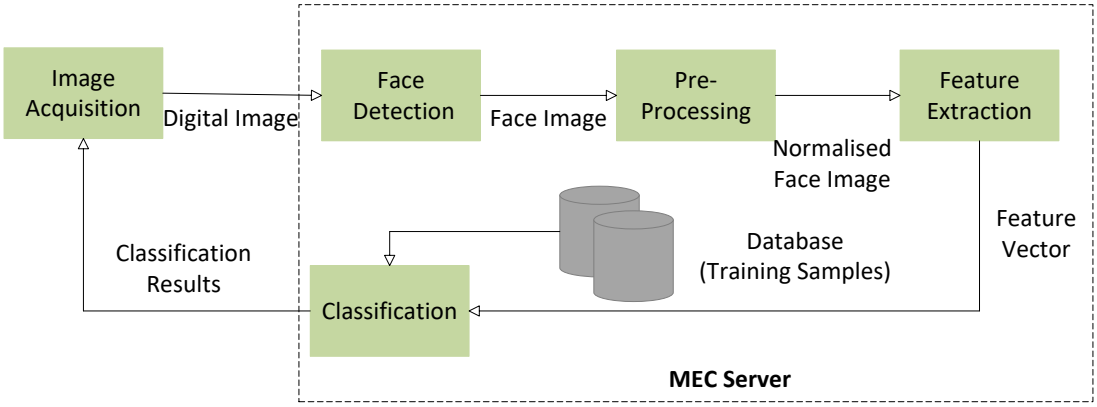


Fig. 2.4: Main computation components in a face recognition application [4].

of emerging mobile applications that will benefit from MEC, by offloading their computation-intensive tasks to the MEC servers. In the following, two examples are provided to illustrate the basic principles of MEC. One example is the face recognition application, as illustrated in Fig. 2.4, which typically consists of five key computational components, including image acquisition, face detection, pre-processing, feature extraction, and classification [4]. While the image acquisition component must be performed on the mobile device to support the user interface, the remaining components—such as signal processing and Machine Learning (ML) algorithms, which require more complex computation—can be offloaded to the MEC server for processing.

Another widely used application that can benefit from the rich resources at the network edge is Augmented Reality (AR). AR applications combine computer-generated data with the physical world, creating an enhanced interactive experience. AR applications as shown in Fig. 2.5 have five critical components [5, 110], namely, the video source (which obtains raw video frames from the mobile camera), a tracker (which tracks the position of the user), a mapper (which builds a model of the environment), an object recogniser (which identifies known objects in the environment), and a renderer (which prepares the processed frame for display).

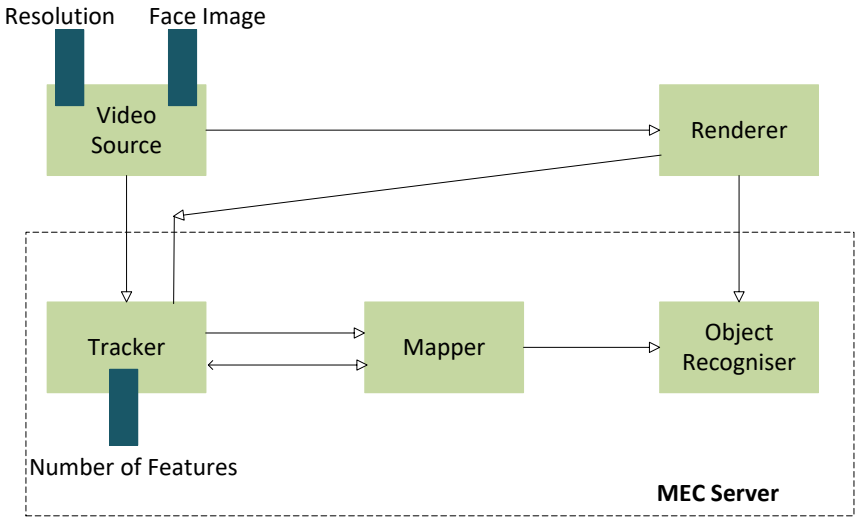


Fig. 2.5: Main computation components in an AR application [5].

The components that require the greatest processing, such as the tracker, mapper and renderer need to be offloaded and executed remotely while computational-non-intensive, video source and renderer, can be executed locally. Thus, mobile users might profit in a number of ways from MEC, including latency decrease and energy saving.

As a key enabling technology for realising the IoT and 5G visions [109, 111], the system models of the typical MEC system for the key computation/communication components are introduced in the following.

The model of MEC is generally divided into four parts, including computation task models, communication models, computation models of mobile devices and computation models of MEC server. There are two computation-task models popularly used in the existing literature on MEC, corresponding to binary and partial offloading, respectively. Binary offloading is known as when a highly integrated or relatively simple task cannot be partitioned and has to be executed as a whole either locally at the mobile device or offloaded to the MEC server. On the contrary, when the program can be partitioned into two parts with one executed at

the mobile device and the remaining offloaded for edge execution, this is known as partial offloading. The generally used communication models are wireless channels. Tasks at the mobile devices and the MEC server are computed by their Central Processing Unit (CPU), respectively. Designing effective MEC systems that optimise both computation offloading and radio resource management is crucial due to the random variations of wireless channels in time, frequency, and space [112–114].

By offloading from mobile devices to MEC servers, the performance of system can improve greatly. In order to achieve high SE and provide QoS for MEC networks, wireless resources must be used effectively and appropriate MA strategies must be implemented [115]. [116–118] have investigated the resource allocation problem in TDMA, FDMA and OFDMA aided-MEC systems, respectively, demonstrating the benefit of offloading. Compared to these OMA schemes, NOMA has been recognised as an enabler to improve both the SE and EE [32, 119–121].

NOMA has been widely investigated and applied in the MEC network to improve EE, reliability and/or latency [122–132]. The authors combined NOMA and MEC in [122–124, 129, 130] to optimise the weighted sum-energy consumption with the constraints of computation latency. Offloading delay minimisation problem of NOMA-MEC network has been investigated in [125, 126], where [125] derived closed-form expressions for the optimal task partition ratio and offloading transmit power and [126] has proposed and compared two algorithms to solve the optimisation problem and further established the criteria to select the working modes of MEC offloading. As SCP which refers to the probability that a computation-intensive task is successfully completed, considering constraints such as latency and resources, is an important metric in a MEC system, an SCP maximisation problem has been formulated and closed-form expressions for SCP have been derived in [127], demonstrating the gain of NOMA over OMA. The overall error probability of NOMA-aided MEC network with the consideration of imperfect SIC has been

studied in [128]. The numerical results have shown the performance improved with the implementation of NOMA, however, multi-antenna NOMA can impose heavy computational burdens on the transmitters and the receivers. Besides, it can also be sensitive to the user deployment. These disadvantages make NOMA not suitable for 6G MEC network [8].

As downlink RSMA bridges and outperforms SDMA, NOMA, and PHY-layer multicasting in multi-antenna networks across various network loads and user deployments [36, 38], it has emerged as a promising MA technique for multi-antenna networks in scenarios with IFBL [31]. There have been a few works investigating the application of RSMA in MEC network [133–135]. The downlink RSMA principles have been utilised to facilitate the offloading of users' computation tasks to multiple MEC servers concurrently [133]. A hybrid RSMA-TDMA scheme for a MEC system has been proposed to minimise delay [134]. The sum of users' maximum delay has been minimised by jointly optimising computation task assignment ratios, transmission power and computational resources. Instead of investigating delay minimisation, the authors in [135] have utilised RSMA in MEC system to achieve the maximum achievable rate of secondary users while maintaining the performance of primary users. Besides, the authors have formulated an SCP maximisation problem and derived the closed-form expressions for SCP. With the aid of RSMA, the performance of MEC network has been enhanced compared to NOMA-aided MEC system [127].

According to the latency-critical requirement of MEC, investigating MEC in FBL regime becomes necessary. Some research has been carried out on MEC in FBL regime [128, 136–138] in terms of EE, reliability, and latency, respectively. However, all those works are based on TDMA or NOMA transmission schemes. Therefore, the analysis of RSMA-aided MEC system in the FBL regime remains a research problem.

Chapter 3

Uplink SISO RSMA in Finite Blocklength Regime

This chapter investigates the performance of RSMA with FBL constraints in short-packet uplink communications, focusing on two crucial aspects: sum-throughput and rate region maximisation. With the consideration of a two-user uplink SISO system, optimal power and rate allocation factors are designed to maximise the sum-throughput and rate region maximisation, exploring the impact of blocklength and target rate on the throughput and error probability performance of RSMA where one user message is split. The superiority of uplink SISO RSMA in the FBL regime is demonstrated via simulations in terms of sum-throughput and rate region, i.e., RSMA is capable of achieving a higher sum-throughput with the same transmission rates, and a larger rate region with the same error probabilities constraint over conventional schemes, respectively. The study concludes that uplink RSMA can outperform NOMA in terms of throughput and error probability performance.

3.1 Introduction

Since the low-latency requirements in URLLC force systems to operate with short blocklength, analysing the system performance with FBL codes is considered to be necessary [6, 7], bringing out a stringent latency requirement to the physical layer. Under the assumption of IFBL, uplink RSMA has been demonstrated that it can provide significant gains in terms of outage probability, sum-throughput, sum-rate performance, high minimum rate, and lower latency compared to NOMA and OMA in [51]. Motivated by the benefits of RSMA in the IFBL regime, RSMA is introduced into FBL uplink communications.

This chapter, motivated by exploring the benefits of uplink RSMA in IFBL regime, alongside the identified research gaps, investigates sum-throughput regarding error probability and rate-region performance of uplink RSMA with FBL constraints in a two-user system for the first time. The main contributions are summarised as follows:

- First, for the *first time*, RSMA is introduced as a powerful multiple access scheme optimised for uplink multi-antenna networks in the context of short-packet communications. This work stands out for its practical consideration of transmitting data packets with finite blocklength, a critical aspect for supporting low-latency applications in 5G and future wireless networks.
- Second, the performance of uplink RSMA under FBL constraints is investigated. This pioneering endeavour aims to optimise RSMA in the FBL regime, tackling both sum-throughput maximisation and rate-region optimisation challenges. Based on the established model, the analytical expressions of error probability for each user are derived. The formulated optimisation problems under FBL constraints are addressed by the proposed SCA-based method. The power and rate allocations among the splitting messages for one

user are designed for two problems with different constraints, i.e. fixed transmission rate for sum-throughput maximisation and fixed error probability for rate-region maximisation.

- Through simulations, the performance of uplink RSMA in a AWGN channel is analysed. In the context of sum-throughput maximisation, the results demonstrate that RSMA outperforms NOMA by achieving a higher throughput with the same transmission rates with and without the aid of time-sharing. Moreover, RSMA can achieve the same throughput as NOMA with a shorter blocklength, thereby reducing the latency. As for the rate-region maximisation problem, the numerical results show that RSMA is capable of achieving a larger rate-region than NOMA with the same error probability constraint. Extensive results demonstrate uplink RSMA achieves more reliable communication with FBL codes compared to uplink NOMA by splitting the message of one user into two parts.

3.2 System Model

3.2.1 RSMA

In this subsection, a two-user uplink scenario with perfect CSIT and CSIR is considered, where two single-antenna users indexed by $\mathcal{K} = \{1, 2\}$ communicate with a single antenna BS. Following the uplink RSMA principle in [51], one user's message is split, resulting in a total of three streams transmitted to the BS. Fig. 3.1 illustrates the two-user uplink RSMA example for the SISO channel. User-1 splits its message W_1 into two parts, $W_{1,i}, i \in (1, 2)$, and W_2 denote the message of user-2. These three messages $W_{1,1}$, $W_{1,2}$, and W_2 are encoded into streams $s_{1,1}$, $s_{1,2}$ and s_2,

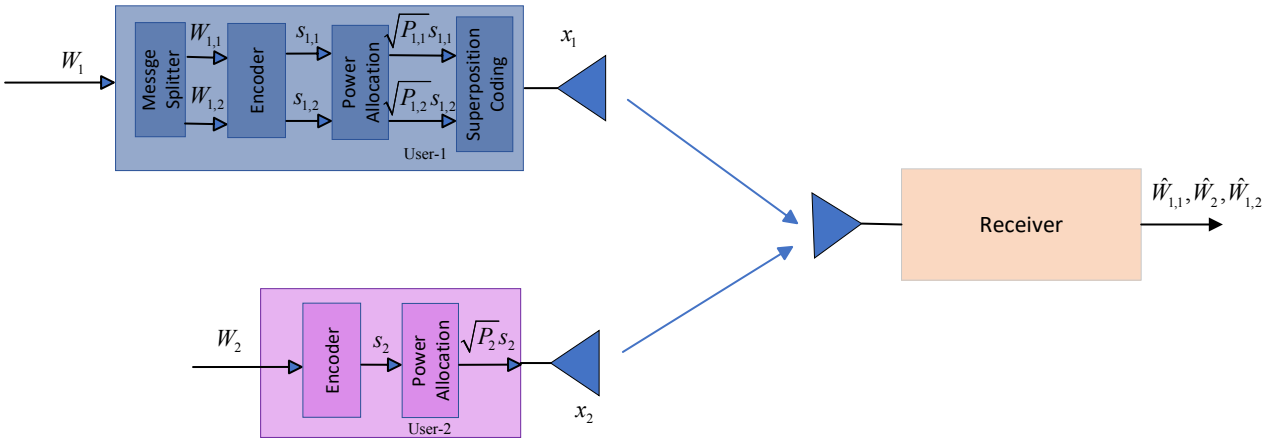


Fig. 3.1: Uplink SISO RSMA for the two-user system.

respectively. The received signal is expressed as

$$y = h_1\sqrt{P_{1,1}}s_{1,1} + h_1\sqrt{P_{1,2}}s_{1,2} + h_2\sqrt{P_2}s_2 + n, \tag{3.1}$$

where $P_{1,i}$ is the transmit power of $s_{1,i}, i \in (1,2)$, respectively and P_2 is the transmit power of s_2. They are limited by the power budget that $\sum_{i=1}^{2} P_{1,i} \leq P_t$ and $P_2 \leq P_t$. The scalar $n \sim \mathcal{CN}(0, \sigma_n^2)$ is the AWGN at the BS. All possible decoding orders at the BS can be classified into three cases, i.e., (i) $s_{1,1} \to s_2 \to s_{1,2}$ (i.e. the BS decodes $s_{1,1}$ first, followed by s_2, and finally $s_{1,2}$), (ii) $s_{1,1} \to s_{1,2} \to s_2$ and (iii) $s_2 \to s_{1,1} \to s_{1,2}$[1]. The decoding order (i) is chosen to write the SINR sequentially.

Accordingly, $s_{1,1}$ is decoded first at the BS by treating other signals as noise. The SINR $\gamma_{1,1}^1$ of the first decoded stream $s_{1,1}$ is

$$\gamma_{1,1}^1 = \frac{P_{1,1}|h_1|^2}{P_{1,2}|h_1|^2 + P_2|h_2|^2 + \sigma_n^2}, \tag{3.2}$$

where superscript 1 represents the index of the splitting user. Assuming that $s_{1,1}$ is successfully decoded, it is reconstructed into $\widehat{W}_{1,1}$ and subtracted from the original

[1]By exchanging the orders of $s_{1,1}$ and $s_{1,2}$, there are three other decoding orders which are ignored in this chapter without loss of generality.

received signal y to obtain y'. The SINR of the second decoded stream s_2 is

$$\gamma_2^1 = \frac{P_2|h_2|^2}{P_{1,2}|h_1|^2 + \sigma_n^2}. \tag{3.3}$$

Assuming that s_2 is successfully decoded to obtain $\widehat{W}_2$, it is reconstructed and subtracted from the received signal y'. Assuming $s_{1,1}$ and s_2 are correctly decoded, the SINR of the last decoded stream $s_{1,2}$ is

$$\gamma_{1,2}^1 = \frac{P_{1,2}|h_1|^2}{\sigma_n^2}. \tag{3.4}$$

Finally, if $s_{1,2}$ is successfully decoded, the estimated message $\widehat{W}_1$ for user-1 is obtained by combining $\widehat{W}_{1,1}$ and $\widehat{W}_{1,2}$.

3.2.2 NOMA

NOMA is a particular instance of the system model of Section 3.2.1, since $0 \leq P_{1,1} \leq P_1$, NOMA can be regarded as a subset of RSMA. The messages W_1 and W_2 from user-1 and user-2 are encoded into s_1 and s_2, respectively. The received signal at the BS is expressed as

$$y = h_1\sqrt{P_1}s_1 + h_2\sqrt{P_2}s_2 + n. \tag{3.5}$$

The decoding order 'NOMA-1' is assumed as $s_1 \rightarrow s_2$ to demonstrate the SINR sequentially. Accordingly, the SINRs of streams under the assumption of successful

decoding are written as

$$\gamma_1^{N1} = \frac{P_1|h_1|^2}{P_2|h_2|^2 + \sigma_n^2} \tag{3.6a}$$

$$\gamma_2^{N1} = \frac{P_2|h_2|^2}{\sigma_n^2}, \tag{3.6b}$$

where the superscript 'N1' stands for 'NOMA-1'.

3.3 Error Probability Analysis

The finite blocklength achievable rate expression is given by [61]

$$R \approx \frac{1}{2}\log_2(1+\gamma) - \sqrt{\frac{V}{N}}Q^{-1}(\epsilon), \tag{3.7}$$

where $V \triangleq V(\gamma) = \log_2^2(e)\left(1 - (1+\gamma)^{-2}\right)$ is the channel dispersion, ϵ is the error probability, γ is the SINR of the stream, N is the blocklength. Based on (3.7), the error probability for a given transmission rate can be written as

$$\epsilon = Q\left(\frac{\frac{1}{2}\log_2(1+\gamma) - R}{\sqrt{V(\gamma)/N}}\right) \triangleq Q(f(\gamma, R)), \tag{3.8}$$

where $f(\gamma, R) \triangleq \left(\frac{1}{2}\log_2(1+\gamma) - R\right)/\sqrt{V(\gamma)/N}$.

3.3.1 RSMA

The target rates of user-1 and user-2 are denoted as R_1 and R_2, respectively. Given that user-1's message is split, the target rates of two split streams are given by $R_{1,1} = \beta R_1$ and $R_{1,2} = (1-\beta)R_1$, where $0 \leq \beta \leq 1$ is the rate allocation factor.

The error probability for each event listed above can be written as

$$\epsilon_a^1 = Q(f(\gamma_{1,1}^1, \beta R_1)) \tag{3.9a}$$

$$\epsilon_b^1 = Q(f(\gamma_2^1, R_2)) \tag{3.9b}$$

$$\epsilon_c^1 = Q(f(\gamma_{1,2}^1, (1-\beta)R_1)), \tag{3.9c}$$

For decoding order (i), the decoding error occurs due to three events listed as[2]: a) $s_{1,1}$ is incorrectly decoded; b) $s_{1,1}$ is correctly decoded but s_2 is incorrectly decoded; c) $s_{1,1}$ and s_2 are both correctly decoded but $s_{1,2}$ is incorrectly decoded.

Thus, the overall error probabilities for the message of user-1 and user-2 can be calculated as

$$\epsilon_1^1 = \epsilon_a^1 + (1-\epsilon_a^1)\epsilon_b^1 + (1-\epsilon_a^1)(1-\epsilon_b^1)\epsilon_c^1 \tag{3.10a}$$

$$\epsilon_2^1 = \epsilon_a^1 + (1-\epsilon_a^1)\epsilon_b^1. \tag{3.10b}$$

3.3.2 NOMA

There are two events for incorrect decoding of s_2 : a) s_1 is incorrectly decoded; b) s_1 is correctly decoded but s_2 is incorrectly decoded. The probabilities of the specified events are calculated by

$$\epsilon_a^{N1} = Q(f(\gamma_1^{N1}, R_1)) \tag{3.11a}$$

$$\epsilon_b^{N1} = Q(f(\gamma_2^{N1}, R_2)). \tag{3.11b}$$

Accordingly, the overall error probabilities for the message of user-1 and user-2 are

[2]Recall that for $\widehat{W_1}$ to be equal to the original message W_1, both $s_{1,1}$ and $s_{1,2}$ have to be correctly decoded.

given by

$$\epsilon_1^{N1} = \epsilon_a^{N1} \tag{3.12a}$$

$$\epsilon_2^{N1} = \epsilon_a^{N1} + (1 - \epsilon_a^{N1})\epsilon_b^{N1}. \tag{3.12b}$$

3.4 Problem Formulation and Algorithm

3.4.1 Sum-throughput Maximisation

According to Shannon capacity definition, any rate inside the capacity region can be achieved with an arbitrarily small error probability under the assumption of IFBL while any rate inside the capacity region cannot be achieved with an arbitrarily small error probability in the FBL regime is possible. Therefore the sum-throughput achieved by RSMA and NOMA is investigated to examine whether RSMA can achieve a lower error probability than NOMA. The throughput of each user is defined as $T_k = R_k\,(1 - \epsilon_k)$ where ϵ_k is the overall error probability of user-k, $k \in \{1, 2\}$. For a given target rate pair (R_k, R_j), an optimisation problem is formulated to maximise the sum-throughput in this subsection as

$$\max_{\mathbf{m},\beta} \quad T_{tot} \tag{3.13a}$$

$$\text{s.t.} \quad P_{k,1} + P_{k,2} \leq P_t,\ k \in \mathcal{K} \tag{3.13b}$$

$$P_j \leq P_t,\ j \neq k \in \mathcal{K} \tag{3.13c}$$

$$0 \leq \beta \leq 1, \tag{3.13d}$$

where $\mathbf{m} = [P_{k,1}, P_{k,2}, P_j]^T$ represents the transmit power of each stream and $T_{tot} = T_k + T_j$ is the sum-throughput derived later in (3.14). The value of β is optimised by exhaustive search and a SCA-based algorithm is proposed to optimise the transmit

power.

Note that for Problem (3.13), a scenario where the message of user-k is split is considered. The transmit power of $s_{k,1}$ and $s_{k,2}$ is represented by $P_{k,1}$ and $P_{k,2}$, respectively and P_j represents the transmit power of user-j ($j \neq k$), which does not perform message splitting. The proposed algorithm is demonstrated over decoding order (i), written as: $s_{k,1} \rightarrow s_j \rightarrow s_{k,2}$. The sum-throughout is defined as

$$T_{tot} = T_k + T_j = R_k + R_j - TP, \tag{3.14}$$

where TP can be calculated as according to the error probability expressions

$$TP = R_k\epsilon_k + R_j\epsilon_j \tag{3.15a}$$

$$= (R_k + R_j)(\epsilon_a^k + \epsilon_b^k - \epsilon_a^k\epsilon_b^k) + R_k\epsilon_c^k(1 - \epsilon_b^k)(1 - \epsilon_a^k). \tag{3.15b}$$

Since $R_k + R_j$ in (3.14) is a constant and the term $\epsilon_a^k\epsilon_b^k\epsilon_c^k$ in (3.15b) so significantly small that can be ignored in the following analysis and minimise the value of TP. Then TP is rewritten as

$$TP = (R_k + R_j)(\epsilon_a^k + \epsilon_b^k) + R_k\epsilon_c^k - (R_k + R_j)(\epsilon_a^k\epsilon_b^k) - R_k(\epsilon_b^k\epsilon_c^k) - R_k(\epsilon_a^k\epsilon_c^k). \tag{3.16}$$

Thus, Problem (3.13) translates into

$$\min_{\mathbf{m},\beta} \quad TP \tag{3.17a}$$

$$\text{s.t.} \quad P_{k,1} + P_{k,2} \leq P_t, \ k \in \mathcal{K} \tag{3.17b}$$

$$P_j \leq P_t, \ j \neq k \in \mathcal{K} \tag{3.17c}$$

$$0 \leq \beta \leq 1. \tag{3.17d}$$

Problem (3.17) is still not convex due to the non-convex multiplied terms $\epsilon_a^k\epsilon_b^k$, $\epsilon_a^k\epsilon_c^k$ and $\epsilon_b^k\epsilon_c^k$ in (3.16). Other slack variables t, $\boldsymbol{\theta} = [\theta_a, \theta_b, \theta_c]^T$ and $\boldsymbol{\rho} = [\rho_{k,1}, \rho_j, \rho_{k,2}]^T$ are introduced. With the aid of new variables, Problem (3.17) is transformed into Problem (3.18).

$$\min_{\mathbf{m},\boldsymbol{\rho},\beta} \quad t \tag{3.18a}$$

$$\text{s.t.} \quad (R_k + R_j)(\epsilon_a^k + \epsilon_b^k) + R_k\epsilon_c^k - (R_k + R_j)(\theta_a^k\theta_b^k) - R_k(\theta_b^k\theta_c^k) - R_k(\theta_a^k\theta_c^k) \leq t,$$

$$k \in \mathcal{K},\ j \neq k \in \mathcal{K} \tag{3.18b}$$

$$\epsilon_i^k \geq \theta_i,\ i \in (a, b, c),\ k \in \mathcal{K} \tag{3.18c}$$

$$\frac{P_{k,1}|h_k|^2}{P_{k,2}|h_k|^2 + P_j|h_j|^2 + \sigma_n^2} \geq \rho_{k,1},\ k \in \mathcal{K},\ j \neq k \in \mathcal{K} \tag{3.18d}$$

$$\frac{P_j|h_j|^2}{P_{k,2}|h_k|^2 + \sigma_n^2} \geq \rho_j,\ k \in \mathcal{K},\ j \neq k \in \mathcal{K} \tag{3.18e}$$

$$\frac{P_{k,2}|h_k|^2}{\sigma_n^2} \geq \rho_{k,2}, k \in \mathcal{K} \tag{3.18f}$$

$$P_{k,1} + P_{k,2} \leq P_t,\ k \in \mathcal{K} \tag{3.18g}$$

$$P_j \leq P_t,\ j \neq k \in \mathcal{K} \tag{3.18h}$$

$$0 \leq \beta \leq 1, \tag{3.18i}$$

Problem (3.18) is still not convex because $\theta_a\theta_b, \theta_a\theta_c$ and $\theta_b\theta_c$ in constraint (3.18b). Therefore, they are approximated at the point $\left(\boldsymbol{\theta}^{[n]}\right)$ at iteration n by the first-order Taylor approximation, which can be written as

$$\theta_a\theta_b \geq \theta_a\theta_b^{[n]} + (\theta_b - \theta_b^{[n]})\theta_a^{[n]} \triangleq \Psi^{[n]}(\theta_a, \theta_b) \tag{3.19a}$$

$$\theta_a\theta_c \geq \theta_a\theta_c^{[n]} + (\theta_c - \theta_c^{[n]})\theta_a^{[n]} \triangleq \Psi^{[n]}(\theta_a, \theta_c) \tag{3.19b}$$

$$\theta_b\theta_c \geq \theta_b\theta_c^{[n]} + (\theta_c - \theta_c^{[n]})\theta_b^{[n]} \triangleq \Psi^{[n]}(\theta_b, \theta_c) \tag{3.19c}$$

For constraint (3.18c), it is also approximated around the point $\left(\boldsymbol{\rho}^{[n]}\right)$ by the first-

order Taylor approximation and ϵ_a^k is taken for example which is given by

$$\epsilon_a^k \geq Q(f(\rho_{k,1}^{[n]}, \beta R_1)) + (\rho_{k,1} - \rho_{k,1}^{[n]}) \frac{dQ(f(\rho_a^{[n]}, \beta R_1))}{d\rho_a^{[n]}} \triangleq \Phi^{[n]}(\rho_{k,1}). \tag{3.20}$$

To keep $\epsilon_a^k \geq \Phi^{[n]}(\rho_{k,1})$, an additional constraint is defined as

$$\log_2(1 + \rho_{k,1}) - \beta R_k \geq 0, \ k \in \mathcal{K}. \tag{3.21}$$

Therefore, constraint (3.18c) is approximated around the point $\left(\boldsymbol{\rho}^{[n]}\right)$ at iteration n as

$$\begin{aligned}
\Phi^{[n]}(\rho_{k,1}) &\geq \theta_a \\
\Phi^{[n]}(\rho_j) &\geq \theta_b \\
\Phi^{[n]}(\rho_{k,1}) &\geq \theta_c \\
\log_2(1 + \rho_{k,1}) - \beta R_k &\geq 0, \ k \in \mathcal{K} \\
\log_2(1 + \rho_j) - R_j &\geq 0, \ j \neq k \in \mathcal{K} \\
\log_2(1 + \rho_{k,2}) - (1 - \beta) R_k &\geq 0, \ k \in \mathcal{K}.
\end{aligned} \tag{3.22}$$

Constraints (3.18d) and (3.18e) are respectively approximated around the point $\left(\mathbf{m}^{[n]}, \boldsymbol{\rho}^{[n]}\right)$ at iteration n by

$$\begin{aligned}
&P_{k,2}|h_k|^2 + P_j|h_j|^2 + \sigma_n^2 - \frac{P_{k,1}|h_k|^2}{\rho_{k,1}^{[n]}} + (\rho_{k,1} - \rho_{k,1}^{[n]}) \frac{P_{k,1}^{[n]}|h_k|^2}{(\rho_{k,1}^{[n]})^2} \leq 0, \ k \in \mathcal{K} \\
&P_{k,2}|h_k|^2 + \sigma_n^2 - \frac{P_j|h_j|^2}{\rho_j^{[n]}} + (\rho_j - \rho_j^{[n]}) \frac{P_j^{[n]}|h_j|^2}{(\rho_j^{[n]})^2} \leq 0, \ j \neq k \in \mathcal{K}.
\end{aligned} \tag{3.23}$$

Based on the above approximation methods, the original non-convex problem is transformed into a convex one and can be solved using the SCA method. SCA solves the problem by approximating a sequence of convex sub-problems. At iteration n,

based on the optimal solution $\left(\mathbf{m}^{[n-1]}, \boldsymbol{\theta}^{[n-1]}, \boldsymbol{\rho}^{[n-1]}\right)$ obtained from the previous iteration $n-1$, the following problem is solved:

$$\min_{\mathbf{m},\boldsymbol{\theta},\boldsymbol{\rho},\beta} \quad t \tag{3.24a}$$

$$\text{s.t.} \quad (R_k + R_j)(\epsilon_a + \epsilon_b) + R_k\epsilon_c - (R_k + R_j)\Psi^{[n]}(\theta_a, \theta_b) - R_k\Psi^{[n]}(\theta_b, \theta_c)$$

$$- R_k\Psi^{[n]}(\theta_a, \theta_c) \leq t, \ k \in \mathcal{K}, \ j \neq k \in \mathcal{K}, \tag{3.24b}$$

$$(3.18\text{f}), (3.18\text{g}), (3.18\text{h}), (3.18\text{i}), (3.22), (3.23). \tag{3.24c}$$

The SCA-based power allocation algorithm is outlined in Algorithm 1 and τ is the tolerance of convergence. Since the solution of Problem (3.24) at iteration $n-1$ is a feasible point of Problem (3.24) at iteration n and power is constrained by (3.18g) and (3.18h), the objective function t is monotonically increasing which implies that the convergence of this proposed SCA-based algorithm is guaranteed. This algorithm for this optimisation problem has two parts which are optimal rate allocation factor β found by exhaustive search-based algorithm and optimal power allocation $P_{k,1}$ and $P_{k,2}$ found by SCA-based algorithm solved in CVX in MATLAB. The computational complexity of the exhaustive search-based algorithm is $\mathcal{O}(\delta^{-1})$ where $\delta \in (0, 1)$ is the increment between two adjacent candidates of β. The candidate β is assumed to be uniformly distributed over the range $(0, 1]$. The SCA-based algorithm in CVX can be solved by interior-point methods with computational complexity $\mathcal{O}\left([K+1]^{3.5}\right)$ [139, 140]. The total number of iterations required for

Algorithm 1: Proposed SCA-based Algorithm for solving Problem (3.24)

Initialise: $n \leftarrow 0, t^{[n]} \leftarrow 0, \mathbf{m}^{[n]}, \boldsymbol{\theta}^{[n]}, \boldsymbol{\rho}^{[n]}$;

repeat

  $n \leftarrow n + 1$;

  Solve problem (3.24) using $\mathbf{m}^{[n-1]}, \boldsymbol{\theta}^{[n-1]}, \boldsymbol{\rho}^{[n-1]}$ and find optimal t^*, $\mathbf{m}^*, \boldsymbol{\theta}^*, \boldsymbol{\rho}^*$;

  Update $t^{[n]} \leftarrow t^*, \mathbf{m}^{[n]} \leftarrow \mathbf{m}^*, \boldsymbol{\theta}^{[n]} \leftarrow \boldsymbol{\theta}^*, \boldsymbol{\rho}^{[n]} \leftarrow \boldsymbol{\rho}^*$;

until $|t^{[n]} - t^{[n-1]}| \leq \tau$;

the convergence is approximated as $\mathcal{O}\left(\log(\tau^{-1})\right)$. The parameter τ is the convergence tolerance of the SCA-based algorithm to update the variables. Hence, the worst-case computational complexity is $\mathcal{O}\left(\delta^{-1}\log(\tau^{-1})[K+1]^{3.5}\right)$.

3.4.2 Rate-region Maximisation

In this subsection, the objective is to maximise the rate region achieved by NOMA and RSMA without considering time-sharing by optimising the power allocation of the splitting messages with fixed error probability $\epsilon = 10^{-3}$. The problem is formulated as

$$\max_{\mathbf{m}} \quad \mu_k R_k + \mu_j R_j \tag{3.25a}$$

$$\text{s.t.} \quad P_{k,1} + P_{k,2} \leq P_t,\ k \in \mathcal{K} \tag{3.25b}$$

$$P_j \leq P_t,\ j \neq k \in \mathcal{K}, \tag{3.25c}$$

where $\mathbf{m} = [P_{k,1}, P_{k,2}, P_j]^T$ represents the transmit power of each stream, and μ_k and μ_j are the weights for user-k and user-j, respectively and $\mu_k + \mu_j = 1$. Since the rate and SINR equations are not convex, the proposed SCA-based algorithm in Sec. 3.4.1 is still applied to solve this problem by introducing slack variables and successively approximating the non-convex expressions by their first-order Taylor approximation. Specifically, slack variables $\mathbf{a} = [a_{k,1}, a_j, a_{k,2}]^T$ and $\boldsymbol{\rho} = [\rho_{k,1}, \rho_j, \rho_{k,2}]^T$

are first introduced, Problem (3.25) is transformed into

$$\max_{\mathbf{m},\boldsymbol{\rho},\mathbf{a}} \quad \mu_k(a_{k,1}+a_{k,2})+\mu_j a_j \tag{3.26a}$$

$$\text{s.t.} \quad \log_2(1+\rho_{k,i})-\frac{Q^{-1}(\epsilon)}{\sqrt{N}}\sqrt{V_{\rho_{k,i}}}\geq a_{k,i},\ k\in\mathcal{K},\ i\in\{1,2\} \tag{3.26b}$$

$$\log_2(1+\rho_j)-\frac{Q^{-1}(\epsilon)}{\sqrt{N}}\sqrt{V_{\rho_j}}\geq a_j,\ j\neq k\in\mathcal{K} \tag{3.26c}$$

$$\frac{P_{k,1}|h_k|^2}{P_{k,2}|h_k|^2+P_j|h_j|^2+\sigma_n^2}\geq\rho_{k,1},\ k\in\mathcal{K}, \tag{3.26d}$$

$$\frac{P_j|h_j|^2}{P_{k,2}|h_k|^2+\sigma_n^2}\geq\rho_j,\ j\neq k\in\mathcal{K} \tag{3.26e}$$

$$\frac{P_{k,2}|h_k|^2}{\sigma_n^2}\geq\rho_{k,2},\ k\in\mathcal{K} \tag{3.26f}$$

$$(3.25\text{b}),(3.25\text{c}).$$

Problem (3.26) remains non-convex due to the constraints (3.26b)-(3.26e). For constraint (3.26d)-(3.26e), they are approximated around the point $\left(\mathbf{m}^{[n]},\boldsymbol{\rho}^{[n]}\right)$ at iteration n by (3.23) in Sec. 3.4.1. Next, the non-convex terms $\sqrt{V_{\rho_{k,i}}}$ and $\sqrt{V_{\rho_j}}$ in constraints (3.26b) and (3.26c) are approximated by their first-order Taylor approximation around the point $\left(\boldsymbol{\rho}^{[n]}\right)$ at iteration n, which are expressed as

$$\begin{aligned}\sqrt{V_{\rho_{k,i}}}&\leq\sqrt{1-(1+\rho_{k,i}^{[n]})^{-2}}+(\rho_{k,i}-\rho_{k,i}^{[n]})(1+\rho_{k,i}^{[n]})^{-3}(1-(1+\rho_{k,i}^{[n]})^{-2})^{-\frac{1}{2}}\\&\triangleq\Phi_{k,i}^{[n]},\ k\in\mathcal{K},\ i\in\{1,2\}\end{aligned} \tag{3.27a}$$

$$\begin{aligned}\sqrt{V_{\rho_j}}&\leq\sqrt{1-(1+\rho_j^{[n]})^{-2}}+(\rho_j-\rho_j^{[n]})(1+\rho_j^{[n]})^{-3}(1-(1+\rho_j^{[n]})^{-2})^{-\frac{1}{2}}\\&\triangleq\Phi_j^{[n]},\ j\neq k\in\mathcal{K}.\end{aligned} \tag{3.27b}$$

Based on the above approximation methods, the original non-convex problem is transformed into a convex one and can be solved using the SCA method. At iteration n, based on the optimal solution $\left(\mathbf{m}^{[n-1]},\boldsymbol{\rho}^{[n-1]}\right)$ obtained from the previous

iteration $n-1$, the following problem is solved:

$$\max_{\mathbf{m},\boldsymbol{\rho},\mathbf{a}} \quad R_{\text{sum}} \tag{3.28a}$$

$$\text{s.t.} \quad \log_2(1+\rho_{k,i}) - \frac{Q^{-1}(\epsilon)}{\sqrt{N}}\Phi_{k,i}^{[n]} \geq a_{k,i},\ k \in \mathcal{K},\ i \in \{1,2\} \tag{3.28b}$$

$$\log_2(1+\rho_j) - \frac{Q^{-1}(\epsilon)}{\sqrt{N}}\Phi_{j}^{[n]} \geq a_j,\ j \neq k \in \mathcal{K} \tag{3.28c}$$

$$(3.23), (3.25\text{b}), (3.25\text{c}),$$

where $R_{\text{sum}} = \mu_k(a_{k,1}+a_{k,2})+\mu_j a_j$ is defined for convenience, and $R_{\text{sum}}^{[n]} = \mu_k(a_{k,1}^{[n]} + a_{k,2}^{[n]}) + \mu_j a_j^{[n]}$. The SCA-based power allocation algorithm is outlined in Algorithm 2. The parameter τ represents the tolerance of convergence.

At each SCA iteration, the approximated problem is solved. Since the solution of Problem (3.28) at iteration $n-1$ is a feasible point of Problem (3.28) at iteration n, and the objective function R_{sum} is monotonically increasing and it is bounded by the transmit power constraints (3.25b) and (3.25c). The total number of SCA iterations required for convergence is approximated as $\mathcal{O}\left(\log(\tau^{-1})\right)$. Hence, the worst-case of computational complexity of the SCA is $\mathcal{O}\left(\log(\tau^{-1})[K+1]^{3.5}\right)$.

Algorithm 2: Proposed SCA-based Algorithm for solving Problem (3.28)

Initialise: $n \leftarrow 0, R_{\text{sum}}^{[n]} \leftarrow 0, \mathbf{m}^{[n]}, \boldsymbol{\rho}^{[n]}, \mathbf{a}^{[n]}$;
repeat
  $n \leftarrow n+1$;
  Solve problem (3.28) using $\mathbf{m}^{[n-1]}, \boldsymbol{\rho}^{[n-1]}, \mathbf{a}^{[n-1]}$ and find optimal R_{sum}^{*}, $\mathbf{m}^{*}, \boldsymbol{\rho}^{*}, \mathbf{a}^{*}$;
  Update $R_{\text{sum}}^{[n]} \leftarrow R_{\text{sum}}^{*}, \mathbf{m}^{[n]} \leftarrow \mathbf{m}^{*}, \boldsymbol{\rho}^{[n]} \leftarrow \boldsymbol{\rho}^{*}, \mathbf{a}^{[n]} \leftarrow \mathbf{a}^{*}$;
until $|R_{sum}^{[n]} - R_{sum}^{[n-1]}| \leq \tau$;

3.5 Numerical Results and Discussion

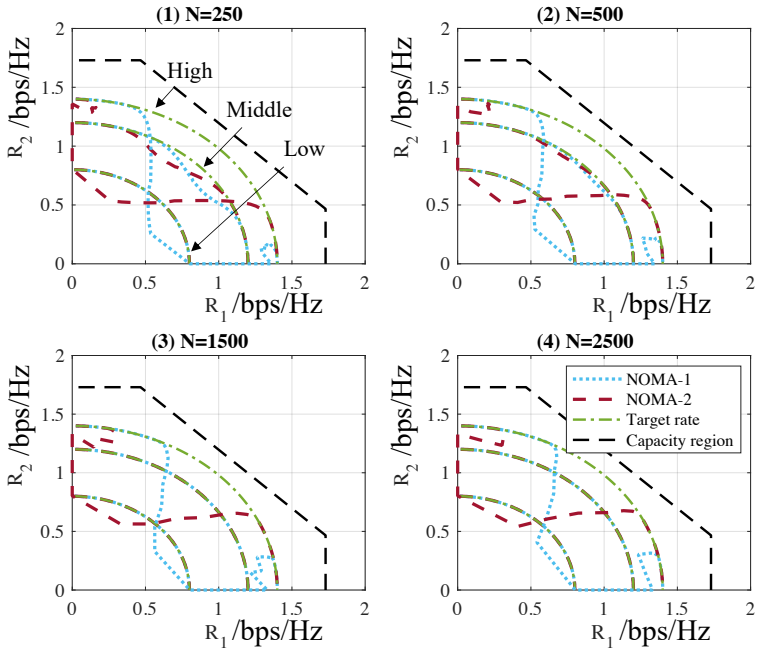
In this section, simulations for the AWGN channel are performed. Without loss of generality, it is assumed that the noise variance is 1 and $P_t = 10$ dB. Results of four different blocklength which are $N = 250, 500, 1500$ and 2500 are compared. To evaluate the sum-throughput achieved by RSMA and NOMA according to different target rate pairs, three kinds of symmetric target rate pairs are chosen which are noted as low, middle and high inside the capacity region with IFBL, respectively. To make target rate pairs symmetric, these three kinds of target rate pairs are set to be on three circles with different radius. Thus, the selected target rate pairs are on the circles with radius of 0.8, 1.2 and 1.4 bps/Hz corresponding to low, middle and high, respectively. Then, the rate region of NOMA and RSMA is compared with different blocklengths with an error probability threshold. 'NOMA-1' and 'NOMA-2' refer to the NOMA with the decoding orders $s_1 \to s_2$ and $s_2 \to s_1$, respectively. 'RSMA-1' and 'RSMA-2' refer to the case where the message of user-1 and user-2 is split, respectively.

3.5.1 Sum-throughput

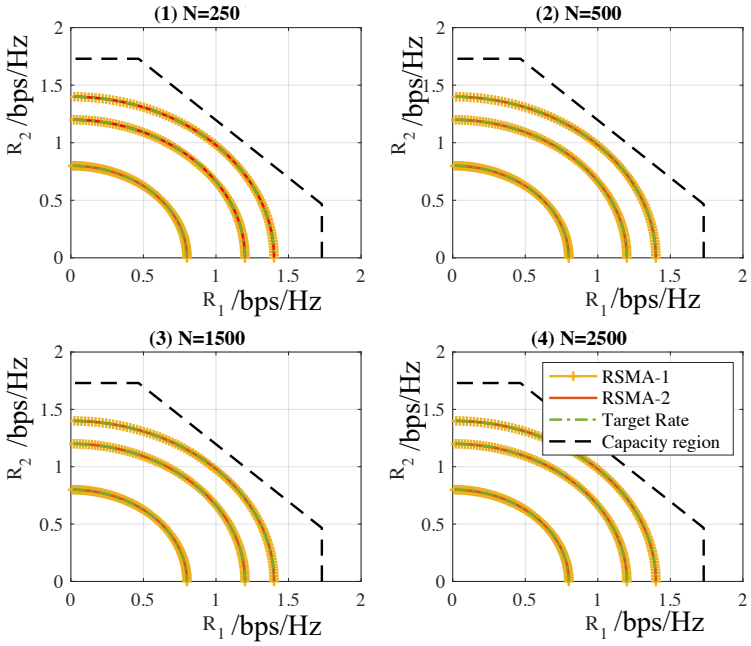
3.5.1.1 Without Time-Sharing

In Fig. 3.2(a), NOMA can achieve the low and middle target rate pairs with 500 blocklength at least. For high rates (consider 'NOMA-1' only), when r_1 is small 'NOMA-1' can achieve them with low error probability although r_2 is high since P_2 can be set high. As r_1 increases the performance deteriorates rapidly especially for user-2 even with a large blocklength. It can be observed that there is a little bump when r_2 approaches 0 in Fig. 3.2(a) since when r_2 is less than a certain value, small P_2 can achieve r_2 and cause little interference to user-1. In Fig. 3.2(b), note

that ‘RSMA-1’ and ‘RSMA-2’ have the same performance and RSMA can achieve the high target rate with very low error probabilities even with short blocklength without time-sharing.



(a) NOMA without time-sharing



(b) RSMA without time-sharing

Fig. 3.2: Throughput of NOMA and RSMA without time-sharing

3.5.1.2 With Time-Sharing

Compared to the result of no time-sharing in Fig. 3.2(a), the performance of NOMA shown in Fig. 3.3 becomes better and can achieve the same throughput as RSMA by transmitting at two target rate pairs with time-sharing. However, RSMA still outperforms NOMA when target rate pairs are high.

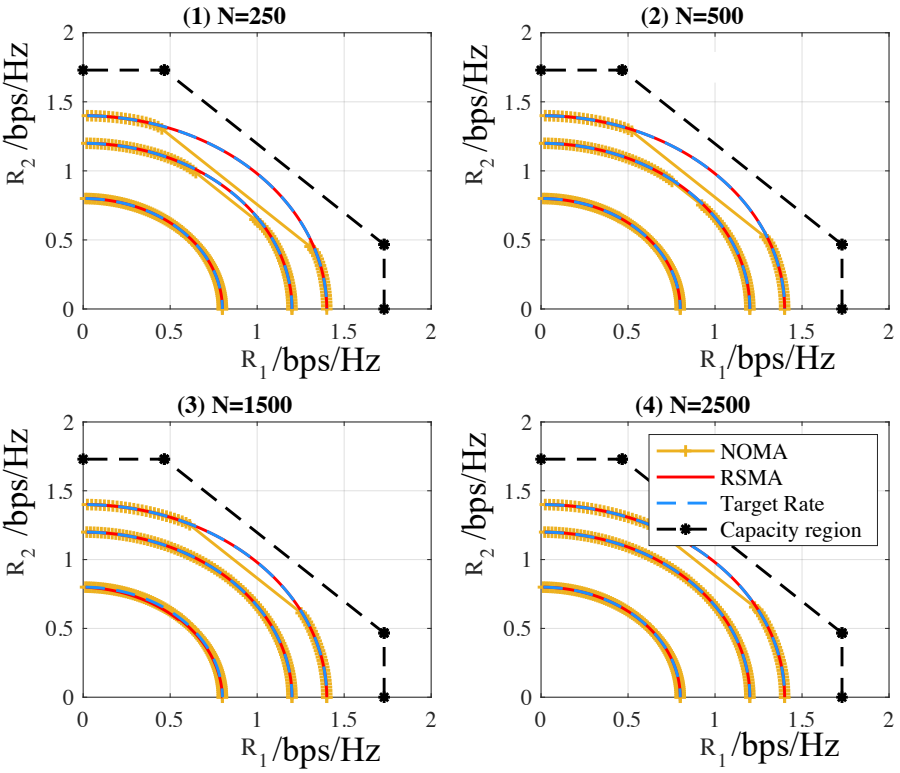


Fig. 3.3: Throughput of NOMA and RSMA with time-sharing.

3.5.2 Rate Region

Since 'NOMA-1' and 'NOMA-2' cannot achieve all the high target rate pairs in Sec. 3.5.1.1, the max target rate pairs that can be achieved by 'NOMA-1' and RSMA are investigated. In the simulation, the error probability is set to $\epsilon = 10^{-3}$ and the rate region achieved by RSMA and 'NOMA-1' with this error probability constraint is plotted.

In Fig. 3.4, one can observe that as the blocklength increases, the rate region of RSMA becomes closer to the capacity region while the rate region of 'NOMA-1' is always smaller than RSMA. It can be concluded that RSMA with decoding order of $s_{1,1} \rightarrow s_2 \rightarrow s_{1,2}$ can achieve a larger rate region while NOMA can only achieve it by cooperating with time-sharing. The reason is that if all the transmit power is allocated to $s_{1,1}$, RSMA performs like 'NOMA-1' and if all the transmit power is allocated to $s_{1,2}$, RSMA performs like 'NOMA-2'. By changing the allocation of the transmit power between two split streams, RSMA bridges 'NOMA-1' and 'NOMA-2' without time-sharing.

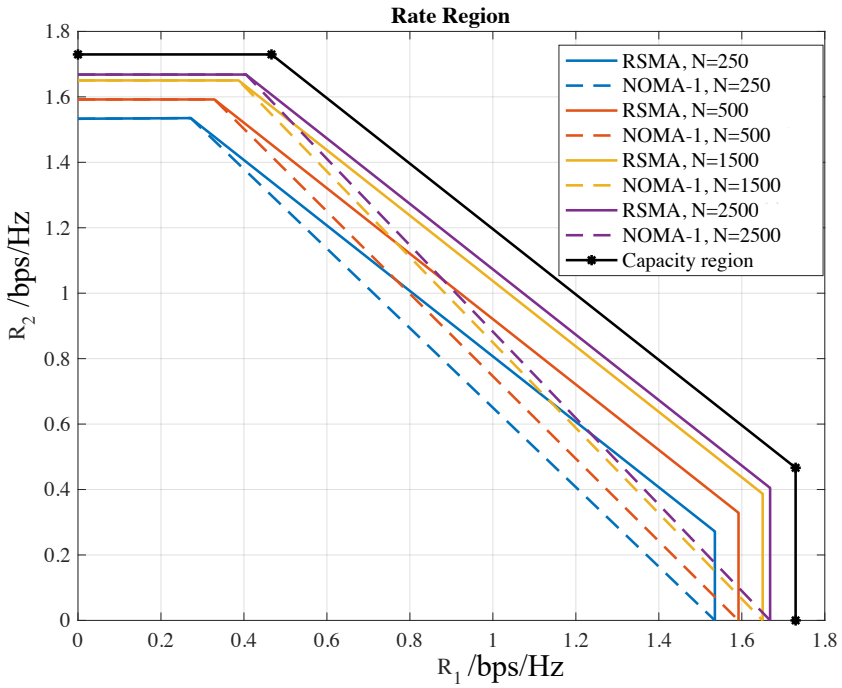


Fig. 3.4: Rate region of two users with $\epsilon = 10^{-3}$.

3.6 Summary

This chapter studies whether RSMA can improve the error probability performance, sum-throughput and the rate region with FBL constraints in uplink SISO channel. The numerical results show that RSMA can achieve significantly higher performance than NOMA without time-sharing. When time-sharing is included, NOMA can achieve the same sum-throughput as RSMA by transmitting at two different target rate pairs while RSMA can directly achieve it by transmitting at one target rate pair. When users transmit at the same target rate pair with RSMA and NOMA respectively, RSMA outperforms NOMA. Secondly, the max transmit target rate with constrained error probabilities is studied. The result shows that the rate region of RSMA is always larger than NOMA without time-sharing. In summary, when the target rates are fixed, RSMA provides lower error probabilities than NOMA and when the error probabilities are fixed, RSMA achieves a larger rate region than NOMA.

Chapter 4

Uplink MIMO RSMA in Finite Blocklength Regime

Given the insights gained from the investigation of uplink SISO RSMA in Chapter 3, this chapter delves into the performance of MIMO RSMA in the context of short-packet uplink communications in FBL regime. Different from Chapter 3 where only two users are included, a K-user system is considered in this chapter and an uplink MIMO RSMA framework is proposed. Both precoders and combiners instead of power allocation (in Chapter 3) are optimised with MMF metric. Due to the high coupling between precoders and combiners, AO is applied to decompose the optimisation problem into two subproblems where each is solved by a SCA-based approach. Additionally, a low-complexity scheme to design the decoding order at the receiver is introduced. The PHY-layer of the uplink MIMO RSMA architecture is designed and evaluated using multi-user LLS. Numerical results demonstrate that uplink MIMO RSMA with FBL constraints can not only achieve MMF gains over conventional transmission schemes such as SDMA and NOMA but also exhibits robustness to network loads. The benefits of splitting messages from multiple users are also illustrated.

4.1 Introduction

The seminal contribution of [61] laid the theoretical foundation for analysing FBL codes in AWGN channels, a framework subsequently extended to MIMO channels [86]. This extension is particularly relevant for the exploration of MIMO systems within the FBL regime. While extensive research has showcased the capacity-enhancing and latency-reducing capabilities of SDMA and NOMA in wireless networks, RSMA emerges as a novel MA technique with the potential to not only bridge but also surpass the performance of SDMA and NOMA.

While downlink RSMA has been shown to provide significant gains in MIMO settings and FBL over SDMA and NOMA [75, 106, 107], there has not been any work on uplink RSMA for MIMO and FBL. Work [141] paved the way towards that goal but is limited to SISO. Therefore, this chapter fills this gap thoroughly by providing theoretical foundations and verifications through LLS. The contributions of the chapter are summarised as follows:

- First, for the *first time*, a general K-user uplink MIMO RSMA system in FBL regime is established. This framework is general in the sense that any arbitrary user can have multiple streams with each being split into two parts and transmitted to the BS. This is the first work investigating the performance of uplink MIMO RSMA systems with FBL constraints.
- Second, based on the established system model, a MMF optimisation problem is formulated with FBL constraints by designing the precoders of the split messages for each user and combiners at the BS. Because of the high coupling between precoders and combiners, AO is utilised by decomposing the original problem into two subproblems which are precoder optimisation and combiner optimisation, respectively. A SCA-based method is proposed to address the formulated subproblems. This is the first work that optimises both precoders

and combiners for uplink MIMO RSMA systems with FBL constraints. Additionally, a low-complexity algorithm is introduced for designing the decoding order at the receiver side.

- With the assumption of perfect CSIT and CSIR, numerical results demonstrate that in the K-user uplink MIMO, the fairness of FBL RSMA with optimised precoders and combiners outperforms NOMA and SDMA schemes even when splitting the messages of only one user. This contrasts with the convention of splitting the messages of $K-1$ users as outlined in [51]. Furthermore, by increasing the number of splitting users, RSMA with the FBL constraint can outperform NOMA with IFBL settings.

- The performance of RSMA in practical systems is illustrated by developing the PHY-layer architecture of RSMA with finite constellation modulation techniques, finite length polar codes, and Adaptive Modulation and Coding (AMC) [73]. Through such LLS, it is demonstrated that RSMA outperforms SDMA and NOMA in uplink MU-MIMO systems in terms of max-min throughput. This is the first study to develop the PHY-layer architecture and assess uplink RSMA performance using LLS.

4.2 System Model

This section firstly models the uplink MIMO RSMA framework, from message splitting strategies at each user's transmitters to the decoding strategy at the BS using SIC. Secondly, the uplink MIMO NOMA is also briefly introduced and explained as a subset of the proposed RSMA scheme. The achievable rates under both MIMO RSMA and NOMA scenarios with the practical FBL consideration are also expressed, to facilitate future optimisation and simulations.

4.2.1 MIMO RSMA

Here the RSMA framework proposed for uplink MIMO channel is presented. A K-user uplink with perfect CSIT and CSIR is consideredwhere users indexed by $\mathcal{K} = \{1, 2, ..., K\}$ are equipped with N_t antennas. It is assumed that the BS has N_r antennas, receiving messages from the K users. The transmitted signal from user-k, $\mathbf{x}_k \in \mathbb{C}^{N_t \times 1}$, $k \in \mathcal{K}$ is subject to a power constraint $\mathbb{E}\{\|\mathbf{x}_k\|^2\} \leq P_t$. The signal is transmitted through a MIMO channel with $\mathbf{H}_k \in \mathbb{C}^{N_r \times N_t}$, $k \in \mathcal{K}$ denoting the channel matrix between user-k and BS. The transmission model of the proposed uplink MIMO with RSMA at the user side is illustrated in Fig. 4.1 and explained below.

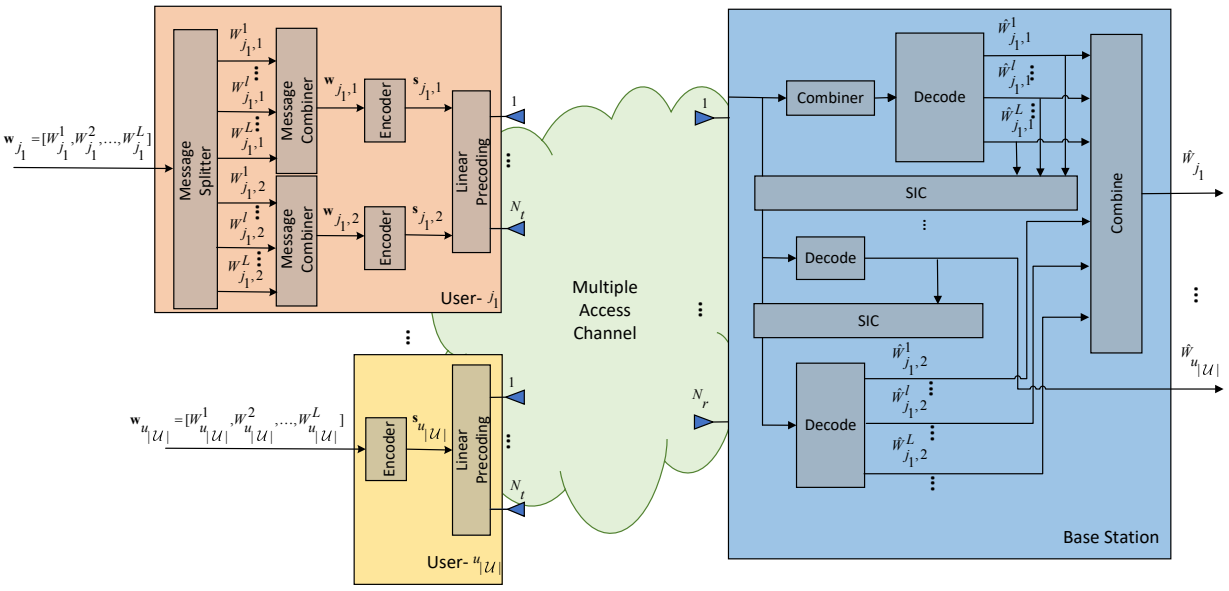


Fig. 4.1: Transmission Model for uplink MU-MIMO adopting RSMA at the user side. Users are classified into splitting users (depicted in orange block) and non-splitting users (depicted in yellow block), with each user in the former group splitting their respective messages to experience different decoding priorities at the BS, leveraging the benefit of rate-splitting.

4.2.1.1 Transmitter

At the user side, let $L = \min\{N_t, N_r\}$ denote the number of message streams for user-k and these messages are expressed as $\mathbf{w}_k = [W_k^1, W_k^2, \ldots, W_k^L]^T,\ k \in \mathcal{K}$. $\mathcal{J} \triangleq \{j_1, \ldots, j_{|\mathcal{J}|}\}$ and $\mathcal{U} = \mathcal{K} \setminus \mathcal{J} \triangleq \{u_1, \ldots, u_{|\mathcal{U}|}\}$ are defined as the splitting users set and the non-splitting users set, respectively. Each message in $\mathbf{w}_{j_l}$ of user-j_l is split into two parts $W_{j_l}^i = \{W_{j_l,1}^i, W_{j_l,2}^i\},\ \forall i \in \{1, \ldots, L\},\ \forall l \in \{1, \ldots, |\mathcal{J}|\},\ j_l \in \mathcal{J}$. The first and second part of the messages for user-j_l are combined together and denoted as $\mathbf{w}_{j_l,1} = [W_{j_l,1}^1, W_{j_l,1}^2, \ldots, W_{j_l,1}^L]^T$ and $\mathbf{w}_{j_l,2} = [W_{j_l,2}^1, W_{j_l,2}^2, \ldots, W_{j_l,2}^L]^T$, respectively. Separating each message into two parts leverages the freedom of varied decoding orders at the receiver. Then they are encoded into two symbol vectors $\mathbf{s}_{j_l,1} = [s_{j_l,1}^1, s_{j_l,1}^2, \ldots, s_{j_l,1}^L]^T \in \mathbb{C}^{L\times 1}$ and $\mathbf{s}_{j_l,2} = [s_{j_l,2}^1, s_{j_l,2}^2, \ldots, s_{j_l,2}^L]^T \in \mathbb{C}^{L\times 1}$, respectively. On the other hand, the messages of non-splitting user-u_q, $\mathbf{w}_{u_q}$ are encoded into $\mathbf{s}_{u_q} = [s_{u_q}^1, s_{u_q}^2, \ldots, s_{u_q}^L]^T \in \mathbb{C}^{L\times 1},\ \forall q \in \{1, \ldots, |\mathcal{U}|\},\ u_q \in \mathcal{U}$. Therefore, by splitting $|\mathcal{J}|$ users' messages, there are $2|\mathcal{J}| + |\mathcal{U}|$ symbol vectors received at the BS in this K-user system, which are combined together into the vector $\mathbf{s} = [\mathbf{s}^{\mathcal{J}}, \mathbf{s}^{\mathcal{U}}]^T$, where $\mathbf{s}^{\mathcal{J}} = [\mathbf{s}_{j_1,1}, \mathbf{s}_{j_1,2}, \ldots, \mathbf{s}_{j_{|\mathcal{J}|},1}, \mathbf{s}_{j_{|\mathcal{J}|},2}]^T$, and $\mathbf{s}^{\mathcal{U}} = [\mathbf{s}_{u_1}, \ldots, \mathbf{s}_{u_{|\mathcal{U}|}}]$, respectively. Linear precoders $\mathbf{P}_{j_l} = [\mathbf{P}_{j_l,1}, \mathbf{P}_{j_l,2}]$ is used to precode the symbol vectors of spliting user-j_l with $tr(\mathbf{P}_{j_l}\mathbf{P}_{j_l}^H) \leq P_t,\ j_l \in \mathcal{J}$. $\mathbf{P}_{u_q}$ is used to precode the symbol vectors of non-spliting user-u_q, $tr(\mathbf{P}_{u_q}\mathbf{P}_{u_q}^H) \leq P_t,\ u_q \in \mathcal{U}$, limited by the power budget P_t at each user. Based on the above illustration, the received signal at the BS, $\mathbf{y} \in \mathbb{C}^{N_r\times 1}$, is given by

$$\mathbf{y} = \underbrace{\sum_{j_l\in\mathcal{J}} \mathbf{H}_{j_l} \sum_{d=1}^{2} \mathbf{P}_{j_l,d}\mathbf{s}_{j_l,d}}_{\text{splitting users}} + \underbrace{\sum_{u_q\in\mathcal{U}} \mathbf{H}_{u_q}\mathbf{P}_{u_q}\mathbf{s}_{u_q}}_{\text{non-splitting users}} + \mathbf{n}, \tag{4.1}$$

where $\mathbf{n} \sim \mathcal{CN}(\mathbf{0}, \sigma_n^2\mathbf{I}_{N_r})$ denotes the AWGN vector at the BS. Without loss of generality, it is assumed that the noise variance is one, i.e., $\sigma_n^2 = 1$. Then, the transmit SNR is numerically equal to the transmit power.

4.2.1.2 Receiver

At the BS side, the signals are received at the receiver and the SIC is employed. There are $(2|\mathcal{J}| + |\mathcal{U}|)!$ possible decoding orders. The optimal decoding order can be determined through exhaustive search, but its complexity grows exponentially with the number of users and data streams. To address this issue, a low-complexity decoding order scheme proposed in the previous work [142] is applied, which involves separating the parts of splitting messages from users in $\mathcal{J}$ and arranging non-splitting users in $\mathcal{U}$ by the descending order of the channel gain. Assume that $\|\mathbf{H}_{u_{j_1}}\| \geq \|\mathbf{H}_{u_{j_2}}\| \geq \ldots \geq \|\mathbf{H}_{u_{j_{|\mathcal{J}|}}}\|$ and $\|\mathbf{H}_{u_{q_1}}\| \geq \|\mathbf{H}_{u_{q_2}}\| \geq \ldots \geq \|\mathbf{H}_{u_{q_{|\mathcal{U}|}}}\|$, the proposed decoding order is $\mathbf{s}_{j_1,1} \rightarrow \cdots \rightarrow \mathbf{s}_{j_{|\mathcal{J}|},1} \rightarrow \mathbf{s}_{u_1} \rightarrow \cdots \rightarrow \mathbf{s}_{u_{|\mathcal{U}|}} \rightarrow \mathbf{s}_{j_1,2} \rightarrow \cdots \rightarrow \mathbf{s}_{j_{|\mathcal{J}|},2}$. This decoding order is mapped into $\boldsymbol{\eta} = [\mathbf{s}_{\eta_1}, ..., \mathbf{s}_{\eta_{2|\mathcal{J}|+|\mathcal{U}|}}]$. These $2|\mathcal{J}| + |\mathcal{U}|$ symbol vectors are indexed by $\mathcal{M} = \{1, ..., 2|\mathcal{J}| + |\mathcal{U}|\}$. Therefore, the m^{th} filtered received signal $\mathbf{z}_{\eta_m},\ m \in \mathcal{M}$ before the m^{th} stage of SIC process is

$$\mathbf{z}_{\eta_m} = \underbrace{\mathbf{G}_{\eta_m}\mathbf{H}_{\eta_m}\mathbf{P}_{\eta_m}\mathbf{s}_{\eta_m}}_{\text{desired signal}} + \underbrace{\sum_{i=m+1}^{2|\mathcal{J}|+|\mathcal{U}|} \mathbf{H}_{\eta_i}\mathbf{P}_{\eta_i}\mathbf{s}_{\eta_i} + \mathbf{G}_{\eta_m}\mathbf{n}}_{\text{interference plus noise}}, \tag{4.2}$$

where $\mathbf{G}_{\eta_m} \in \mathbb{C}^{L \times N_r},\ m \in \mathcal{M}$ is the combiner which is used to detect symbol vector $\mathbf{s}_{\eta_m}$ and η_m refers to the m^{th} symbol vector in the decoding order $\boldsymbol{\eta}$. This m^{th} symbol vector $\mathbf{s}_{\eta_m}$ is decoded at the BS by treating the remaining $m + 1^{\text{th}}$ to $2|\mathcal{J}| + |\mathcal{U}|^{\text{th}}$ symbol vectors as noise. Once $\mathbf{s}_{\eta_m}$ is successfully decoded, $\widehat{\mathbf{w}}_{\eta_m}$ is obtained, reconstructed and subtracted from the current received signal.

4.2.1.3 Rate with FBL constraint

Since the FBL code is considered here, the FBL achievable rate expression derived in [61] is applied. Then the rate of symbol vector $\mathbf{s}_{\eta_m}$ is written as

$$R_{\eta_m} = \sum_{a=1}^{L} \log_2(1+\gamma_{\eta_m}^a) - \frac{B}{\sqrt{N}}\sqrt{V_{\gamma_{\eta_m}^a}}, \tag{4.3}$$

where $\gamma_{\eta_m}^a$ denotes the SINR of stream a of symbol vector $\mathbf{s}_{\eta_m}$

$$\gamma_{\eta_m}^a = \frac{|\mathbf{g}_{\eta_m}^a \mathbf{H}_{\eta_m} \mathbf{p}_{\eta_m}^a|^2}{I_{l,\eta_m}^a + I_{c,\eta_m}^a + \|\mathbf{g}_{\eta_m}^a\|\sigma_n^2}, \tag{4.4}$$

with $\mathbf{p}_{\eta_m}^a$ (resp. $\mathbf{g}_{\eta_m}^a$) the precoder (resp. combiner) attached to stream a of symbol vector $\mathbf{s}_{\eta_m}$ and is the a^{th} column (resp. row) of $\mathbf{P}_{\eta_m}$ (resp. $\mathbf{G}_{\eta_m}$). I_{l,η_m}^a is the inter-stream interference and I_{c,η_m}^a is the interference from other symbol vectors suffered by stream a of symbol vector $\mathbf{s}_{\eta_m}$ which are given as

$$I_{l,\eta_m}^a = \sum_{i=1, i\neq a}^{L} |\mathbf{g}_{\eta_m}^a \mathbf{H}_{\eta_m} \mathbf{p}_{\eta_m}^i|^2 \tag{4.5a}$$

$$I_{c,\eta_m}^a = \sum_{j=m+1}^{2|\mathcal{J}|+|\mathcal{U}|} \sum_{i=1}^{L} |\mathbf{g}_{\eta_m}^a \mathbf{H}_{\eta_j} \mathbf{p}_{\eta_j}^i|^2. \tag{4.5b}$$

As for the penalty part brought by FBL, $B = Q^{-1}(\epsilon)\log_2(e)$, ϵ is the error probability and Q is the Q-function. N is the blocklength and $V = \left(1-(1+\gamma_{\eta_m}^a)^{-2}\right)$ is the channel dispersion. The total achievable rates of splitting users can be calculated by summing the transmission rates for the two corresponding split streams, e.g., the rate of user-j_1 in $\mathcal{J}$ equals $R_{\eta_1} + R_{\eta_{1+K}}$ and the rates of non-splitting users such as user-u_1 in $\mathcal{U}$ equals $R_{\eta_{|\mathcal{J}|+1}}$ individually.

4.2.2 MIMO NOMA

The system model of MIMO NOMA is a particular instance of the model specified in Sec. 4.2.1. By grouping all users into the non-splitting user set, MIMO RSMA boils down to MIMO NOMA. Let $L = \min\{N_t, N_r\}$ denote the number of message streams for user-k and they are expressed as $\mathbf{w}_k = [W_k^1, W_k^2, \ldots, W_k^L]^T,\ k \in \mathcal{K}$ and then are encoded into $\mathbf{s}_k = [s_k^1, s_k^2, \ldots, s_k^L]^T \in \mathbb{C}^{L\times 1}$. $\mathbf{P}_k$ is used to precode the symbol vector of user-k, $tr(\mathbf{P}_k\mathbf{P}_k^H) \leq P_t$. The received signal at the BS, $\mathbf{y} \in \mathbb{C}^{N_r\times 1}$, is given by

$$\mathbf{y} = \sum_{k\in\mathcal{K}} \mathbf{H}_k\mathbf{P}_k\mathbf{s}_k + \mathbf{n}. \tag{4.6}$$

The decoding order is arranged according to the descending order of the channel gain and it is assumed that $\|\mathbf{H}_1\| \geq \|\mathbf{H}_2\| \geq \ldots \geq \|\mathbf{H}_K\|$. The proposed decoding order is $\mathbf{s}_1 \to \mathbf{s}_2 \to \cdots \to \mathbf{s}_{K-1} \to \mathbf{s}_K$. Then the SINR of stream a of symbol vector $\mathbf{s}_k,\ k \in \mathcal{K}$ is written as

$$\gamma_k^a = \frac{|\mathbf{g}_k^a\mathbf{H}_k\mathbf{p}_k^a|^2}{I_{l,k}^a + I_{c,k}^a + \|\mathbf{g}_k^a\|\sigma_n^2}, \tag{4.7}$$

where $\mathbf{p}_k^a$ and $\mathbf{g}_k^a$ is a^{th} column of $\mathbf{P}_k$ and a^{th} row of $\mathbf{G}_k$ of stream a in symbol vector $\mathbf{s}_k$, respectively. $I_{l,k}^a$ and $I_{c,k}^a$ is the inter-stream interference and interference from other symbol vectors experienced by stream a of symbol vector $\mathbf{s}_k$ which are given as

$$I_{l,k}^a = \sum_{i=1, i\neq a}^{L} |\mathbf{g}_k^a\mathbf{H}_k\mathbf{p}_k^i|^2 \tag{4.8a}$$

$$I_{c,k}^a = \sum_{j=k+1}^{2|\mathcal{J}|+|\mathcal{U}|} \sum_{i=1}^{L} |\mathbf{g}_k^a\mathbf{H}_j\mathbf{p}_j^i|^2. \tag{4.8b}$$

Then the FBL achievable rate expression of symbol vector $\mathbf{s}_k$ is written as

$$R_k = \sum_{a=1}^{L} \log_2(1+\gamma_k^a) - \frac{B}{\sqrt{N}}\sqrt{V_{\gamma_k^a}}. \tag{4.9}$$

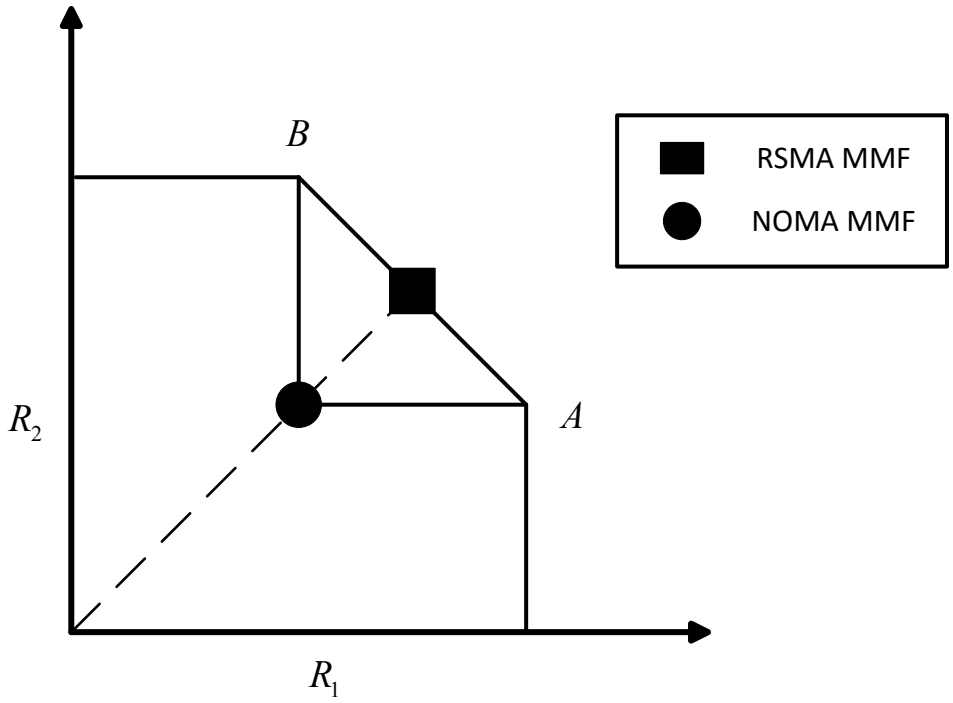


Fig. 4.2: The schematic illustration of the two-user rate-region. The highlighted points (RSMA marked by a square and NOMA marked by a circle) represent the best fairness achieved by RSMA and NOMA respectively.

For simplicity, MIMO RSMA and MIMO NOMA will be referred to as RSMA and NOMA for the rest of the chapter respectively.

Remark: Here, a two-user example is taken to illustrate the benefits of RSMA over NOMA regarding the MMF metric. The capacity region for RSMA and NOMA with random transmit power is illustrated in Fig. 4.2. The capacity region always has the shape of a pentagon. The two corner points, A and B, correspond to the maximum rates achieved by two SIC decoding orders: A is achieved by decoding user-1 first and B is achieved by decoding user-2 first. The points on the segment AB can be achieved with time-sharing between two decoding orders for NOMA. However, RSMA can directly achieve the points between two corner points without time-sharing. Therefore, without considering time-sharing, MMF achieved by RSMA is higher than that of NOMA which is represented by the black square in Fig. 4.2 and MMF of NOMA is denoted by the black circle.

4.3 Problem Formulation and Optimisation

In this section, the optimisation problem is formulated to design the precoders and combiners to facilitate the RSMA performance, with the primary purpose of guaranteeing communication fairness between users. An effective algorithm is developed to solve the problem and provide the numerical complexity analysis for the proposed algorithm.

4.3.1 Problem Formulation

In this section, an optimisation problem is formulated with respect to both precoders and combiners for the proposed system model. The user fairness issue is emphasised in the objective function. Hence, the MMF problem for the K-user uplink RSMA system is formulated as

$$\max_{\mathbf{P}_{j_l},\mathbf{P}_{u_q},\mathbf{G}} \quad \min_{k\in\mathcal{K}} R_k \tag{4.10a}$$

$$\text{s.t.} \quad \text{tr}(\mathbf{P}_{j_l}\mathbf{P}_{j_l}^H) \leq P_t,\ j_l \in \mathcal{J}, \tag{4.10b}$$

$$\text{tr}(\mathbf{P}_{u_q}\mathbf{P}_{u_q}^H) \leq P_t,\ u_q \in \mathcal{U}, \tag{4.10c}$$

where R_k is the rate of user-k in $\mathcal{K}$ and $\mathbf{G}$ is the combiner. It is assumed the maximum transmit power P_t is equal for all users. A SCA-based algorithm is proposed to solve Problem (4.10) because the rate and SINR equations are not convex. This algorithm involves adding slack variables and progressively approximating the non-convex expressions using their first-order Taylor approximation.

First an auxiliary variable t is introduced and for the given decoding order $\boldsymbol{\eta}$,

Problem (4.10) is transferred into

$$\max_{\mathbf{P}_{\boldsymbol{\eta}},\mathbf{G}_{\boldsymbol{\eta}},t} \quad t \tag{4.11a}$$

$$\text{s.t.} \quad R_l + R_{l+K} \geq t,\ l \in \{1,\ldots,|\mathcal{J}|\} \tag{4.11b}$$

$$R_o \geq t,\ o \in \{|\mathcal{J}|+1,\ldots,K\} \tag{4.11c}$$

$$\text{tr}(\mathbf{P}_l\mathbf{P}_l^H) + \text{tr}(\mathbf{P}_{l+K}\mathbf{P}_{l+K}^H) \leq P_t,\ l \in \{1,\ldots,|\mathcal{J}|\} \tag{4.11d}$$

$$\text{tr}(\mathbf{P}_o\mathbf{P}_o^H) \leq P_t,\ o \in \{|\mathcal{J}|+1,\ldots,K\}, \tag{4.11e}$$

where $\mathbf{P}_{\boldsymbol{\eta}} = [\mathbf{P}_{j_1,1},\ldots,\mathbf{P}_{j_{|\mathcal{J}|},1},\mathbf{P}_{u_1},\ldots,\mathbf{P}_{u_{|\mathcal{U}|}},\mathbf{P}_{j_1,2},\ldots,\mathbf{P}_{j_{|\mathcal{J}|},2}]$ and $\mathbf{G}_{\boldsymbol{\eta}} = [\mathbf{G}_{j_1,1},\ldots,\mathbf{G}_{j_{|\mathcal{J}|},1},\mathbf{G}_{u_1},\ldots,\mathbf{G}_{u_{|\mathcal{U}|}},\mathbf{G}_{j_1,2},\ldots,\mathbf{G}_{j_{|\mathcal{J}|},2}]$ are the precoders and combiners of each symbol vector corresponding to the decoding order $\boldsymbol{\eta}$, respectively. (4.11b) and (4.11c) refer to the maximisation of minimum user rate constraints of every splitting user and non-splitting. In the decoding order $\boldsymbol{\eta}$, the symbol vectors $\mathbf{s}_{j_l,1}$ and $\mathbf{s}_{j_l,2}$ are mapped to $\mathbf{s}_l$ and $\mathbf{s}_{l+K}$, $l \in \{1,\ldots,|\mathcal{J}|\}$ and $\mathbf{s}_{u_q}$ is mapped to s_o, $o \in \{|\mathcal{J}|+1,\ldots,K\}$ with both $\{1,\ldots,|\mathcal{J}|\}$ and $\{|\mathcal{J}|+1,\ldots,K\}$ the subsets of $\mathcal{M}$. Slack variables $\boldsymbol{\Omega}_{\boldsymbol{\eta}} = [\boldsymbol{\rho}_1,\ldots,\boldsymbol{\rho}_{2|J|+|\mathcal{U}|}]$ are further introduced where $\boldsymbol{\rho}_m = [\rho_m^1,\ldots,\rho_m^L]^T,\ m \in \mathcal{M}$.

Then Problem (4.11) is transformed into

$$\max_{\mathbf{P}_{\boldsymbol{\eta}},\mathbf{G}_{\boldsymbol{\eta}},\boldsymbol{\Omega}_{\boldsymbol{\eta}},t} \quad t \tag{4.12a}$$

$$\text{s.t.} \quad \sum_{a=1}^{L}\left(\log_2(1+\rho_l^a) - \frac{B}{\sqrt{N}}\sqrt{V_{\rho_l^a}} + \log_2(1+\rho_{l+K}^a) - \frac{B}{\sqrt{N}}\sqrt{V_{\rho_{l+K}^a}}\right) \geq t,$$

$$l \in (1,\ldots,|\mathcal{J}|) \tag{4.12b}$$

$$\sum_{a=1}^{L}\left(\log_2(1+\rho_o^a) - \frac{B}{\sqrt{N}}\sqrt{V_{\rho_o}}\right) \geq t,\ o \in (|\mathcal{J}|+1,\ldots,K) \tag{4.12c}$$

$$\frac{|\mathbf{g}_{\eta_m}^a\mathbf{H}_{\eta_m}\mathbf{p}_{\eta_m}^a|^2}{I_{l,\eta_m}^a + I_{c,\eta_m}^a + \|\mathbf{g}_{\eta_m}^a\|\sigma_n^2} \geq \rho_m^a,\ m \in \mathcal{M}, a \in \{1,\ldots,L\} \tag{4.12d}$$

(4.11d), (4.11e).

However, Problem (4.12) is still non-convex due to the constraint (4.12b) and (4.12c). Thus the non-convex terms $\sqrt{V_{\rho_l^a}}$, $\sqrt{V_{\rho_{l+K}^a}}$ and $\sqrt{V_{\rho_o^a}}$ are approximated by their first-order Taylor approximation around the point $\left(\boldsymbol{\Omega}_{\boldsymbol{\eta}}^{[n]}\right)$ at iteration n, which are expressed as

$$\begin{aligned}
\sqrt{V_{\rho_l^a}} &\leq \sqrt{1-(1+(\rho_l^a)^{[n]})^{-2}} + \left(\rho_l^a - (\rho_l^a)^{[n]}\right)\left(1+(\rho_l^a)^{[n]}\right)^{-3}\left(1-(1+(\rho_l^a)^{[n]})^{-2}\right)^{-\frac{1}{2}} \\
&\triangleq (\Phi_l^a)^{[n]} \qquad (4.13a) \\
\sqrt{V_{\rho_{l+K}^a}} &\leq \sqrt{1-\left(1+(\rho_{l+K}^a)^{[n]}\right)^{-2}} \\
&\quad + \left(\rho_{l+K}^a - (\rho_{l+K}^a)^{[n]}\right)\left(1+(\rho_{l+K}^a)^{[n]}\right)^{-3}\left(1-(1+(\rho_{l+K}^a)^{[n]})^{-2}\right)^{-\frac{1}{2}} \\
&\triangleq \left(\Phi_{l+K}^a\right)^{[n]} \qquad (4.13b) \\
\sqrt{V_{\rho_o}} &\leq \sqrt{1-(1+(\rho_o^a)^{[n]})^{-2}} + (\rho_o^a - (\rho_o^a)^{[n]})\left(1+(\rho_o^a)^{[n]}\right)^{-3}\left(1-(1+(\rho_o^a)^{[n]})^{-2}\right)^{-\frac{1}{2}} \\
&\triangleq (\Phi_o^a)^{[n]}\,. \qquad (4.13c)
\end{aligned}$$

Thus, Problem (4.12) is transformed into

$$\begin{aligned}
&\max_{\mathbf{P}_{\boldsymbol{\eta}},\mathbf{G}_{\boldsymbol{\eta}},\boldsymbol{\Omega}_{\boldsymbol{\eta}},t} \quad t \qquad (4.14a) \\
\text{s.t.} \quad &\sum_{a=1}^{L}\left(\log_2(1+\rho_l^a) - \frac{B}{\sqrt{N}}(\Phi_l^a)^{[n]} + \log_2(1+\rho_{l+K}^a) - \frac{B}{\sqrt{N}}(\Phi_{l+K}^a)^{[n]}\right) \geq t, \\
&l \in (1,\ldots,|\mathcal{J}|) \qquad (4.14b) \\
&\sum_{a=1}^{L}\left(\log_2(1+\rho_o^a) - \frac{B}{\sqrt{N}}(\Phi_o^a)^{[n]}\right) \geq t,\ o \in (|\mathcal{J}|+1,\ldots,K) \qquad (4.14c) \\
&\frac{|\mathbf{g}_{\eta_m}^a \mathbf{H}_{\eta_m} \mathbf{p}_{\eta_m}^a|^2}{I_{l,\eta_m}^a + I_{c,\eta_m}^a + \|\mathbf{g}_{\eta_m}^a\|\sigma_n^2} \geq \rho_m^a,\ m \in \mathcal{M}, a \in \{1,\ldots,L\} \qquad (4.14d) \\
&(4.11\text{d}), (4.11\text{e}).
\end{aligned}$$

Since the precoder $\mathbf{p}_{\eta_m}^a$ and the combiner $\mathbf{g}_{\eta_m}^a$ are coupled in (4.14d), Problem (4.14) is decomposed into the two subproblems of precoders optimisation and combiners optimisation, which can be efficiently solved by the SCA technique as described in

the following.

4.3.2 Precoders Optimisation

Under any given feasible combiner $\mathbf{G}_{\eta_m},\ m \in \mathcal{M}$ with $\mathbf{g}^a_{\eta_m}$ is the combiner for stream a of symbol vector $\mathbf{s}_{\eta_m}$, Problem (4.14) is rewritten as

$$\max_{\mathbf{P}_{\boldsymbol{\eta}},\boldsymbol{\Omega}_{\boldsymbol{\eta}},t} \quad t \tag{4.15a}$$

$$\text{s.t.} \quad \frac{|\mathbf{g}^a_{\eta_m}\mathbf{H}_{\eta_m}\mathbf{p}^a_{\eta_m}|^2}{I^a_{l,\eta_m}+I^a_{c,\eta_m}+\|\mathbf{g}^a_{\eta_m}\|\sigma_n^2} \geq \rho^a_m,\ m \in \mathcal{M}, a \in \{1,\ldots,L\} \tag{4.15b}$$

$$(4.14\text{b}), (4.14\text{c}), (4.11\text{d}), (4.11\text{e}).$$

With fixed combiner $\mathbf{g}^a_{\eta_m}$, constraint (4.15b) can be rewritten as

$$\frac{|\mathbf{g}^a_{\eta_m}\mathbf{H}_{\eta_m}\mathbf{p}^a_{\eta_m}|^2}{\rho^a_m} \geq I^a_{l,\eta_m}+I^a_{c,\eta_m}+\|\mathbf{g}^a_{\eta_m}\|\sigma_n^2. \tag{4.16}$$

Since (4.16) is not convex, or the Light-Hand Side (LHS) of (4.16), the first-order Taylor approximation is used to approximate the LHS of (4.16) around the point $\left(\mathbf{P}^{[n]}_{\boldsymbol{\eta}},\boldsymbol{\Omega}^{[n]}_{\boldsymbol{\eta}}\right)$ at iteration n and it can be expressed as

$$\frac{|\mathbf{g}^a_{\eta_m}\mathbf{H}_{\eta_m}\mathbf{p}^a_{\eta_m}|^2}{\rho^a_m} \geq \frac{2\Re\{\mathbf{g}^a_{\eta_m}\mathbf{H}_{\eta_m}\left((\mathbf{p}^a_{\eta_m})^{[n]}\right)^H(\mathbf{H}_{\eta_m})^H\left(\mathbf{g}^a_{\eta_m}\right)^H\}}{(\rho^a_m)^{[n]}} - \frac{\rho^a_m|\mathbf{g}^a_{\eta_m}\mathbf{H}_{\eta_m}(\mathbf{p}^a_{\eta_m})^{[n]}|^2}{\left((\rho^a_m)^{[n]}\right)^2} \tag{4.17}$$

Then, constraint (4.15b) is rewritten as

$$\begin{aligned} I^a_{l,\eta_m}+I^a_{c,\eta_m}+\|\mathbf{g}^a_{\eta_m}\|\sigma_n^2 &+ \frac{\rho^a_m|\mathbf{g}^a_{\eta_m}\mathbf{H}_{\eta_m}\left(\mathbf{p}^a_{\eta_m}\right)^{[n]}|^2}{\left((\rho^a_m)^{[n]}\right)^2} \\ &- \frac{2\Re\{\mathbf{g}^a_{\eta_m}\mathbf{H}_{\eta_m}\left((\mathbf{p}^a_{\eta_m})^{[n]}\right)^H(\mathbf{H}_{\eta_m})^H\left(\mathbf{g}^a_{\eta_m}\right)^H\}}{(\rho^a_m)^{[n]}} \leq 0. \end{aligned} \tag{4.18}$$

Therefore, with a given combiner $\mathbf{G}_{\boldsymbol{\eta}}$, the Problem (4.15) is transformed into

$$\max_{\mathbf{P}_{\boldsymbol{\eta}},\boldsymbol{\Omega}_{\boldsymbol{\eta}},t} \quad t \tag{4.19a}$$

$$\text{s.t.} \quad (4.14\text{b}), (4.14\text{c}), (4.18), (4.11\text{d}), (4.11\text{e}).$$

Note that the objective value obtained from Problem (4.19) provides a lower bound on that of Problem (4.15) due to the replacement of the lower bounds in (4.17).

4.3.3 Combiners Optimisation

Similar to precoders optimisation, for any given feasible precoder $\mathbf{P}_{\eta_m},\ m \in \mathcal{M}$ where $\mathbf{p}^a_{\eta_m}$ is the precoder of stream a of symbol vector $\mathbf{s}_{\eta_m}$, Problem (4.14) is rewritten as

$$\max_{\mathbf{G}_{\boldsymbol{\eta}},\boldsymbol{\Omega}_{\boldsymbol{\eta}},t} \quad t \tag{4.20a}$$

$$\text{s.t.} \quad (4.14\text{b}), (4.14\text{c}), (4.15\text{b}).$$

It is still the non-convexity of constraint (4.15b) that causes the problem to be hard to solve. By applying the same method used in Sec. 4.3.2, constraint (4.15b) is approximated around the point $\left(\mathbf{G}^{[n]}_{\boldsymbol{\eta}}, \boldsymbol{\Omega}^{[n]}_{\boldsymbol{\eta}}\right)$ at iteration n and it can be written as

$$\begin{aligned} I^a_{l,\eta_m} + I^a_{c,\eta_m} + \|\mathbf{g}^a_{\eta_m}\|\sigma_n^2 + \frac{\rho^a_m |\left(\mathbf{g}^a_{\eta_m}\right)^{[n]} \mathbf{H}_{\eta_m}(\mathbf{p}^a_{\eta_m})|^2}{\left((\rho^a_m)^{[n]}\right)^2} \\ - \frac{2\Re\{\mathbf{g}^a_{\eta_m}\mathbf{H}_{\eta_m}\left(\mathbf{p}^a_{\eta_m}\right)^H (\mathbf{H}_{\eta_m})^H \left((\mathbf{g}^a_{\eta_m})^{[n]}\right)^H\}}{(\rho^a_m)^{[n]}} \leq 0 \end{aligned} \tag{4.21}$$

Therefore, with a given precoder $\mathbf{p}_{\eta_m}^a$, the Problem (4.20) is transformed into

$$\max_{\mathbf{G}_{\boldsymbol{\eta}}, \boldsymbol{\Omega}_{\boldsymbol{\eta}}, t} \quad t \tag{4.22a}$$

$$\text{s.t.} \quad (4.14\text{b}), (4.14\text{c}), (4.21)$$

Similarly, the objective value obtained from Problem (4.22) serves as a lower bound on that of Problem (4.20). Now, it is easy to verify that both Problem (4.19) and (4.22) are convex problems, which can be efficiently solved via standard convex problem solvers such as CVX [143].

4.3.4 Proposed Algorithm, Convergence and Complexity

According to the above two subproblems, AO is applied to solve Problem (4.11) by utilising the SCA-based algorithm. Specifically, the precoder $\mathbf{P}_{\boldsymbol{\eta}}$ and the combiner $\mathbf{G}_{\boldsymbol{\eta}}$ are alternately optimised by solving Problem (4.19) and (4.22). The details of the proposed algorithm are summarised in Algorithm 3. At each iteration of Algorithm 3, the precoders and combiners are updated by solving Second Order Cone Programming (SOCP) problems. Each SOCP is solved by using interior-point method with the computational complexity of $\mathcal{O}\left([X]^{3.5}\right)$, where X is the total number of variables in the corresponding SOCP problem. Although an additional vari-

Algorithm 3: Proposed SCA-based AO algorithm for solving Problem (4.11)

Initialise: $n \leftarrow 0, t^{[n]} \leftarrow 0$, and feasible $\mathbf{P}_{\boldsymbol{\eta}}^{[n]}, \mathbf{G}_{\boldsymbol{\eta}}^{[n]}$;

repeat

- $n \leftarrow n + 1$;
- Find optimal $\mathbf{P}_{\boldsymbol{\eta}}^{[n]}$ by solving Problem (4.19) for given $\mathbf{G}_{\boldsymbol{\eta}}^{[n-1]}$;
- Find optimal $\mathbf{G}_{\boldsymbol{\eta}}^{[n]}$ by solving Problem (4.22) for given $\mathbf{P}_{\boldsymbol{\eta}}^{[n]}$;
- Update $t^{[n]} \leftarrow t^*, \mathbf{P}_{\boldsymbol{\eta}}^{[n]} \leftarrow \mathbf{P}_{\boldsymbol{\eta}}^*, \mathbf{G}_{\boldsymbol{\eta}}^{[n]} \leftarrow \mathbf{G}_{\boldsymbol{\eta}}^*$;

until $|t^{[n]} - t^{[n-1]}| \leq \tau$;

able $\mathbf{\Omega}$ is introduced for approximating convex problems, the main complexity still comes from the precoders and combiners design. With given combiners, the number of variables of Problem (4.19) is given by $X_{\text{precoder}} = (2|\mathcal{J}| + |\mathcal{U}|)\, N_t$. Similarly, the number of variables of Problem (4.22) is given by $X_{\text{combiner}} = (2|\mathcal{J}| + |\mathcal{U}|)\, N_r$. The total number of iterations required for the convergence is $\mathcal{O}\left(\log\left(\tau^{-1}\right)\right)$, where τ is the convergence tolerance of Algorithm 3. Therefore, the total computation complexity of Algorithm 3 is $\mathcal{O}\left(\left((2|\mathcal{J}| + |\mathcal{U}|)\left(N_t + N_r\right)\right)^{3.5} \log\left(\tau^{-1}\right)\right)$.

Next the convergence of Algorithm 3 is demonstrated. Define $t\left(\mathbf{P}_{\boldsymbol{\eta}}^{[n]}, \mathbf{G}_{\boldsymbol{\eta}}^{[n]}\right)$ as the objective value at the n^{th} iteration. First, from Problem (4.19) with a given $\mathbf{G}_{\boldsymbol{\eta}}$ in step 3 of Algorithm 3, it is known that

$$\begin{aligned} t\left(\mathbf{P}_{\boldsymbol{\eta}}^{[n-1]}, \mathbf{G}_{\boldsymbol{\eta}}^{[n-1]}\right) &\overset{a}{=} t_{\mathbf{P}}\left(\mathbf{P}_{\boldsymbol{\eta}}^{[n-1]}, \mathbf{G}_{\boldsymbol{\eta}}^{[n-1]}\right) \\ &\overset{b}{\leq} t_{\mathbf{P}}\left(\mathbf{P}_{\boldsymbol{\eta}}^{[n]}, \mathbf{G}_{\boldsymbol{\eta}}^{[n-1]}\right) \\ &\overset{c}{\leq} t\left(\mathbf{P}_{\boldsymbol{\eta}}^{[n]}, \mathbf{G}_{\boldsymbol{\eta}}^{[n-1]}\right), \end{aligned} \tag{4.23}$$

where $t_{\mathbf{P}}$ represents the objective value of Problem (4.19). a holds since the first-order Taylor approximations are tight at the given point $\left(\mathbf{P}_{\boldsymbol{\eta}}^{[n-1]}, \mathbf{\Omega}_{\boldsymbol{\eta}}^{[n-1]}\right)$. Since the solution of the approximated Problem (4.19) at the $n-1^{\text{th}}$ iteration is a feasible point for Problem (4.19) at the n^{th} iteration, it can be solved successfully. Moreover, the objective function is bounded by the transmit power constraints, then b holds. c is due to the objective value of (4.19) being the lower bound of (4.15).

Similarly, from Problem (4.22) with a given $\mathbf{P}_{\boldsymbol{\eta}}$ in step 4 of Algorithm 3, it is known that

$$\begin{aligned} t\left(\mathbf{P}_{\boldsymbol{\eta}}^{[n]}, \mathbf{G}_{\boldsymbol{\eta}}^{[n-1]}\right) &\overset{a}{=} t_{\mathbf{G}}\left(\mathbf{P}_{\boldsymbol{\eta}}^{[n]}, \mathbf{G}_{\boldsymbol{\eta}}^{[n-1]}\right) \\ &\overset{b}{\leq} t_{\mathbf{G}}\left(\mathbf{P}_{\boldsymbol{\eta}}^{[n]}, \mathbf{G}_{\boldsymbol{\eta}}^{[n]}\right) \\ &\overset{c}{\leq} t\left(\mathbf{P}_{\boldsymbol{\eta}}^{[n]}, \mathbf{G}_{\boldsymbol{\eta}}^{[n]}\right), \end{aligned} \tag{4.24}$$

where $t_{\mathbf{G}}$ represents the objective value of Problem (4.22). a, b and c hold for the same reasons as described previously.

Thus, it can be obtained that

$$t\left(\mathbf{P}_{\boldsymbol{\eta}}^{[n-1]}, \mathbf{G}_{\boldsymbol{\eta}}^{[n-1]}\right) \leq t\left(\mathbf{P}_{\boldsymbol{\eta}}^{[n]}, \mathbf{G}_{\boldsymbol{\eta}}^{[n]}\right), \tag{4.25}$$

which proves that Algorithm 3 generates a non-decreasing sequence of objective values.

4.4 Physical Layer Design

Apart from the theoretical foundations, it is crucial to prove the enhanced efficiency of RSMA in practical scenarios. Practical transceiver architectures for RSMA in downlink MISO and MIMO channels have been designed in [73] and [75]. In this section, a practical transceiver architecture for uplink MIMO RSMA setting is proposed. Fig. 4.3 illustrates the proposed transceiver architecture.

At the transmitter side, finite alphabet modulation schemes 4-QAM, 16-QAM, 64-QAM and 256-QAM, finite-length polar coding [144] for Forward Error Correction (FEC) and an AMC algorithm are employed. The two messages from the splitting user-j_l are mapped to bit vectors $\mathbf{w}_{j_l,d}^{a}$ of length $K_{j_l,d}^{a}$, for $j_l \in \mathcal{J},\ d \in \{1,2\}$ and $a \in \{1,...,L\}$. Similarly, the messages from the non-splitting user-u_q are mapped to bit vectors $\mathbf{w}_{u_q}^{b}$ of length $K_{u_q}^{b}$, for $u_q \in \mathcal{U}$ and $b \in \{1,...,L\}$. Then, the information bit vectors of splitting user-j_l and non-splitting user-u_q are independent and uniformly distributed in $\mathbb{F}_2^{K_{j_l,d}^{a}}$ and $\mathbb{F}_2^{K_{u_q}^{b}}$, respectively. Then the information bit vectors are independently encoded, interleaved and modulated into symbol streams $\mathbf{s}_{j_l,d}^{a}$ and $\mathbf{s}_{u_q}^{b}$, each of length S. The modulation and coding parameters are determined by an AMC algorithm according to the corresponding rate. The precoders for all streams

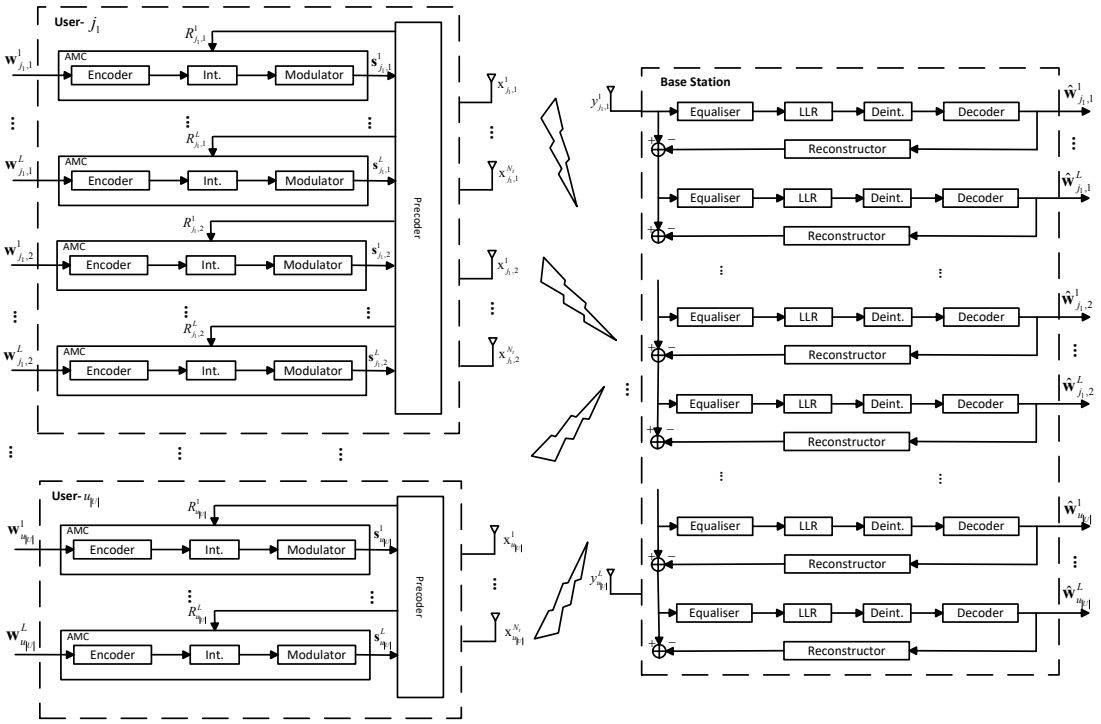


Fig. 4.3: The Proposed Transceiver Architecture of uplink MIMO RSMA for Link-Level Simulations.

are obtained as described in Algorithm 3. More details on the channel coding procedure and AMC algorithm are given in [73] for interested readers.

At the receiver side, instead of performing joint decoding which has high complexity in implementations, interference nulling and interference cancellation are utilised to benefit from low-complexity decoding methods. The received signal is first detected using an equaliser. Minimum Mean Square Error (MMSE) equaliser is applied here. Log-Likelihood Ratios (LLR) are calculated from the combined signal for Soft Decision (SD) decoding of polar codes. Next, the LLRs are de-interleaved and decoded to acquire the decoded bits. The reconstruction process involves repeating operations at the transmitter for the decoded bits to obtain a precoded signal and multiplying this signal with the corresponding channel matrix.

According to the decoding order proposed in Sec. 4.2.1.2, when decoding the first symbol vector $\mathbf{s}_{j_1,1}$, containing L streams, it is assumed that the first $a-1$, $a \leq L$

streams have been correctly decoded and subtracted from the received signal. Then the remaining signal $\tilde{\mathbf{y}}^{\mathbf{s}^a_{j_1,1}}$ can be written as

$$\tilde{\mathbf{y}}^{\mathbf{s}^a_{j_1,1}} = \mathbf{H}_{j_1} \sum_{i=a}^{L} \mathbf{p}^i_{j_1,1} \mathbf{s}^i_{j_1,1} + \mathbf{H}_{j_1} \mathbf{P}_{j_1,2} \mathbf{s}_{j_1,2} + \sum_{j_l, l\neq 1 \in \mathcal{J}} \mathbf{H}_{j_l} \sum_{d=1}^{2} \mathbf{P}_{j_l,d} \mathbf{s}_{j_l,d} + \sum_{u_q \in \mathcal{U}} \mathbf{H}_{u_q} \mathbf{P}_{u_q} \mathbf{s}_{u_q} + \mathbf{n}, \tag{4.26}$$

where the superscript of $\tilde{\mathbf{y}}^{\mathbf{s}^a_{j_1,1}}$ represent the interferences from first $a-1$, $a \leq L$ streams of $\mathbf{s}_{j_1,1}$ have been canceled, $\mathbf{p}^i_{j_1,1} \in \mathbb{C}^{N_t \times 1}$ is the precoder of i^{th} stream of $\mathbf{s}_{j_1,1}$, corresponding the i^{th} column of precoder $\mathbf{P}_{j_1,1}$ and $\mathbf{s}^i_{j_1,1}$ represents the stream i of symbol vector $\mathbf{s}_{j_1,1}$. The corresponding MMSE combiner is given by

$$\mathbf{g}^{\mathbf{s}^a_{j_1,1}} = \left(\mathbf{H}_{j_1} \mathbf{p}^a_{j_1,1}\right)^H \left(\sum_{i=a}^{L} \mathbf{H}_{j_1} \mathbf{p}^i_{j_1,1} {\mathbf{p}^i_{j_1,1}}^H \mathbf{H}^H_{j_1} + \mathbf{N}_{j_1} \right)^{-1}, \tag{4.27}$$

where

$$\mathbf{N}_{j_1} = \mathbf{H}_{j_1} \mathbf{P}_{j_1,2} \mathbf{P}^H_{j_1,2} \mathbf{H}^H_{j_1} + \sum_{j_l, l\neq 1 \in \mathcal{J}} \sum_{d=1}^{2} \mathbf{H}_{j_l} \mathbf{P}_{j_l,d} \mathbf{P}^H_{j_l,d} \mathbf{H}^H_{j_l} + \sum_{u_q \in \mathcal{U}} \mathbf{H}_{u_q} \mathbf{P}_{u_q} \mathbf{P}^H_{u_q} \mathbf{H}^H_{u_q} + \mathbf{I}. \tag{4.28}$$

After combining the signal $\mathbf{g}^{\mathbf{s}^a_{j_1,1}} \tilde{\mathbf{y}}^{\mathbf{s}^a_{j_1,1}}$, LLRs are calculated. The LLRs calculation method is used in [145]. Let $\lambda_i^{\mathbf{s}^a_{j_1}}$ denote the i^{th} bit of the combined stream $\mathbf{s}^a_{j_1,1}$. It is written as

$$\lambda_i^{\mathbf{s}^a_{j_1,1}} = \gamma^{\mathbf{s}^a_{j_1,1}} [\min_{a \in \theta^i_1} \psi(a) - \min_{a \in \theta^i_0} \psi(a)], \tag{4.29}$$

where $\gamma^{\mathbf{s}^a_{j_1,1}}$ is calculated by (4.4) with $m=1$, θ^i_b is the set of modulation symbols with the value b, $b \in \{0,1\}$ at the i^{th} bit location, $\psi(a) = |\frac{\mathbf{g}^{\mathbf{s}^a_{j_1,1}} \tilde{\mathbf{y}}^{\mathbf{s}^a_{j_1,1}}}{\phi_{\mathbf{s}^a_{j_1,1}}} - a|^2$ and $\phi_{\mathbf{s}^a_{j_1,1}} = \frac{\gamma^{\mathbf{s}^a_{j_1,1}}}{1+\gamma^{\mathbf{s}^a_{j_1,1}}}$. A similar procedure is applied to all the remaining streams. The decoding operation utilises a Successive Cancellation List (SCL) decoder for point-to-point polar codes [146].

4.5 Numerical Results and Disucssion

The performance of FBL RSMA with perfect CSIT and CSIR is evaluated and compared with conventional transmission schemes in this section. The MMF performance of uplink MIMO RSMA is illustrated and followed by the LLS results. Here, the performance of RSMA is compared to NOMA and SDMA.

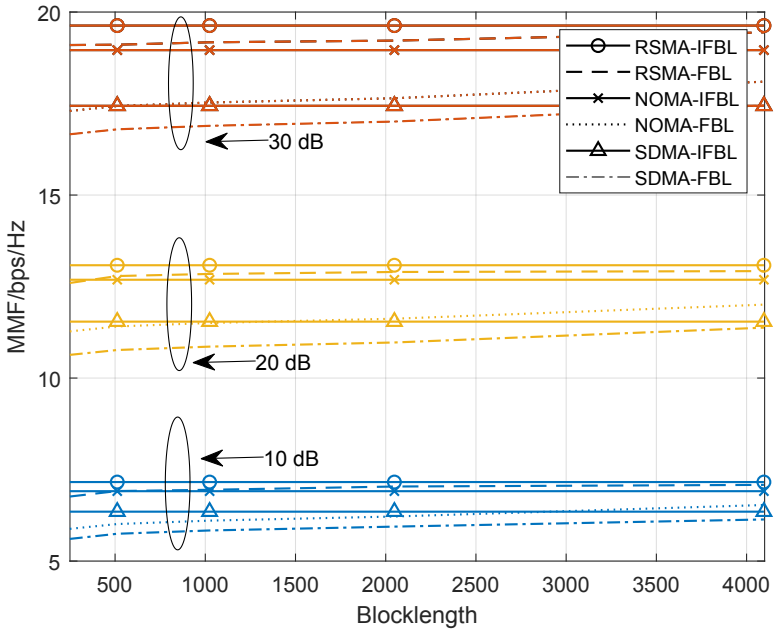
- NOMA: This is a special case of RSMA. As specified in Sec. 4.2.2, none of the users split their messages in NOMA. The BS sequentially decodes the user messages based on SIC. It is applied by expanding the precoder technique suggested in [142] to the MIMO scenario with combiner optimisation mentioned in Sec. 4.3.2.
- SDMA: According to [38], SDMA is a classical method where none of the users split their messages, and the BS decodes the message of each user by treating the messages of all other users as interference.

Simulations for the Rayleigh Fading channel with 100 channel realisations are performed. The tolerance of the algorithm is set to be $\tau = 10^{-4}$.

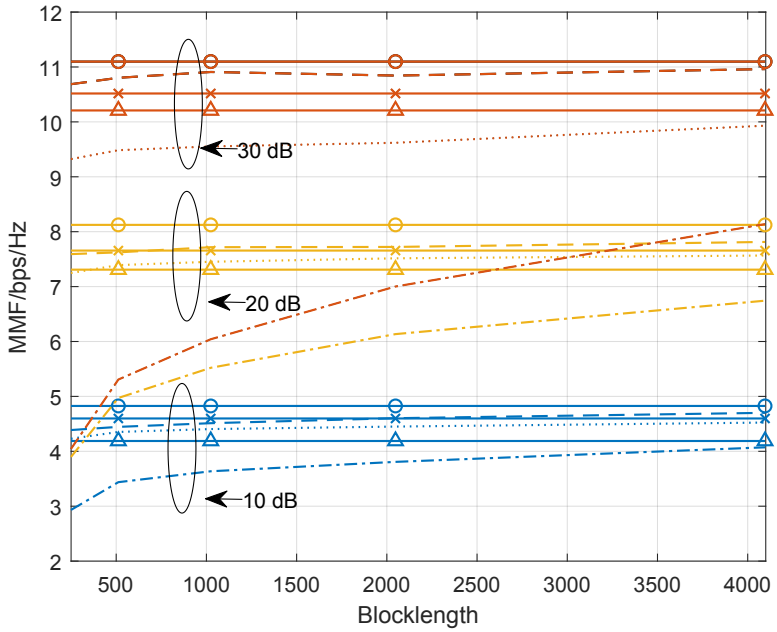
4.5.1 MMF Performance

The trend of MMF as blocklength increases with $K = 2$ is shown in Fig. 4.4. In this simulation, the number of transmit antenna N_t is 2. Blue, yellow and red lines represent the transmit SNR of 10 dB, 20 dB and 30 dB, respectively. Fig. 4.4(a) shows the results of $N_r = 4$, corresponding to the underloaded scenario while the overloaded scenario with $N_r = 2$ is shown in Fig. 4.4(b).

It is evident that in both cases, FBL RSMA consistently outperforms FBL NOMA and FBL SDMA. With underloaded network settings in Fig. 4.4(a), the advantages



(a) $N_r = 4$



(b) $N_r = 2$

Fig. 4.4: The MMF performance of RSMA, NOMA and SDMA versus blocklengths with three different transmit SNR averaged over 100 random channel realisations. $K = 2$. $N_t = 2$. (a) is the underloaded scenario; (b) is the overloaded scenario.

of FBL RSMA become evident when the transmit SNR is high, i.e., FBL RSMA with blocklength larger than 500 even outperforms IFBL NOMA with a transmit

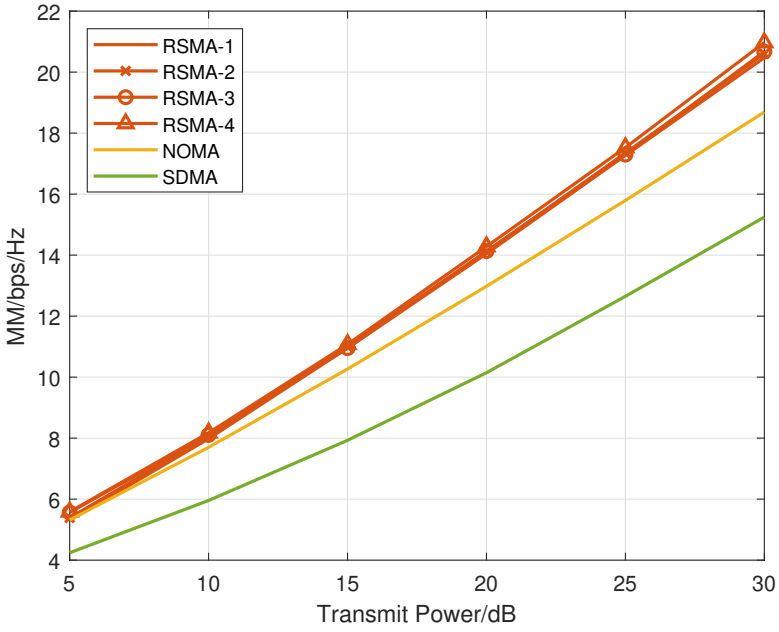
SNR of 30 dB.

In overloaded user settings in Fig. 4.4(b), the performance gain of FBL RSMA over FBL SDMA becomes more significant as the transmit SNR increases due to its adeptness in managing overloaded scenarios. As evidence, given a transmit SNR of 30 dB, FBL SDMA with a 2000 blocklength (the red dashed curve with no marker) is even outperformed by FBL RSMA with a lower transmit SNR (20 dB) and a shorter FBL blocklength (e.g. 1500) (the yellow dashed curve with circle marker). Besides comparison with FBL SDMA, the MMF rate for FBL RSMA with blocklengths of 1500 and 1000 surpasses that of IFBL NOMA when the transmit SNR is 10 dB and 20 dB, respectively. Since the decoding order strictly limits NOMA, the streams of the second decoded user would influence the SINR of the first decoded user significantly. However, with the aid of RSMA, the decoding order can be more flexible to balance the SINR between each decoded stream.

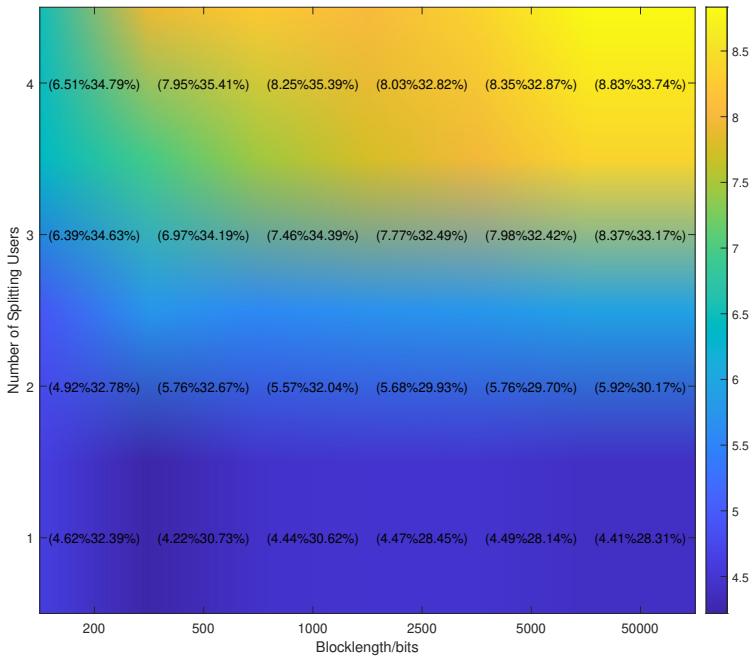
In Fig. 4.4, only splitting one user in the two-user scenario is taken into account as it has been proved that splitting both users has the same effect as splitting one user [56]. Now the effect of the number of splitting users when $K = 4$ is exmained. The MMF of NOMA and SDMA is compared in Fig. 4.5(a) and Fig. 4.6(a) to: i) RSMA with one splitting user; ii) RSMA with two splitting users; iii) RSMA with three splitting users, and iv) RSMA with four splitting users in both underloaded and overloaded cases with FBL constraints. The blocklength is 250. The relative gain of FBL RSMA compared to FBL NOMA and FBL SDMA is defined by (4.30) as

$$\frac{R^{\text{RSMA}} - R^{\text{NOMA/SDMA}}}{R^{\text{NOMA/SDMA}}} \times 100\%. \tag{4.30}$$

Fig. 4.5(b) and Fig. 4.6(b) illustrate the gain of FBL RSMA over FBL NOMA at transmit SNR of 20 dB. The percentages in parentheses represent MMF rate gains of FBL RSMA over FBL NOMA and FBL SDMA (with the same blocklength).

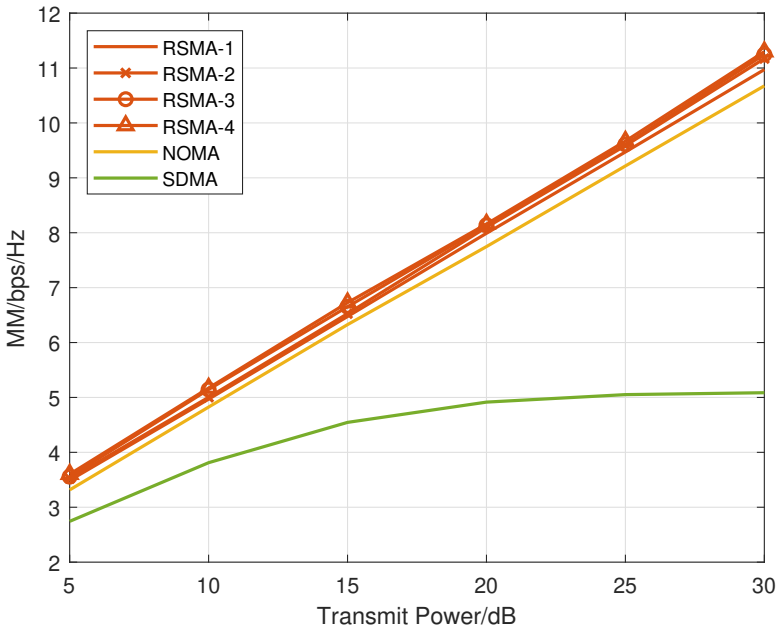


(a) $N_r = 8$

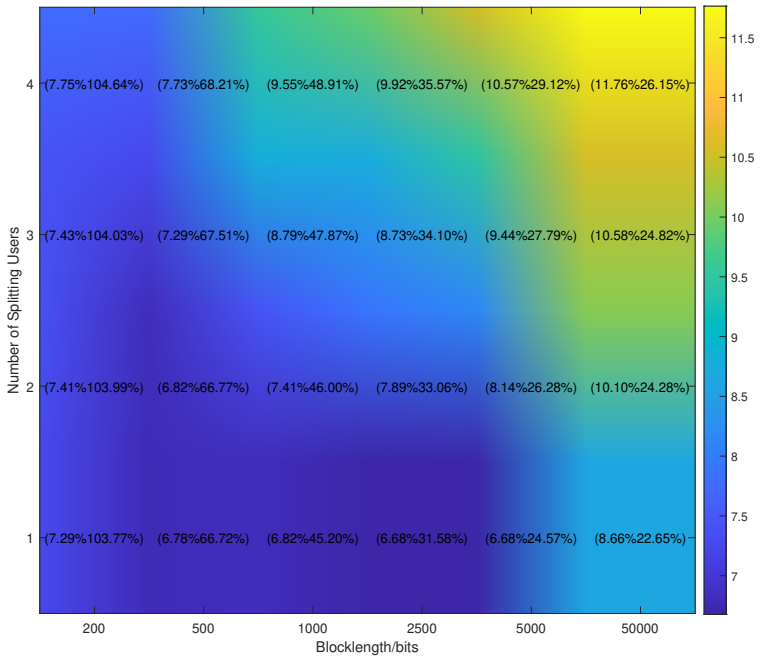


(b) $N_r = 8$

Fig. 4.5: The MMF performance of RSMA, NOMA and SDMA in the underloaded network settings averaged over 100 random channel realisations. Blocklength is 250. $K = 4$. $N_t = 2$. (a) is the MMF versus transmit SNR; (b) is the relative gain of RSMA at the transmit SNR of 20 dB.



(a) $N_r = 4$



(b) $N_r = 4$

Fig. 4.6: The MMF performance of RSMA, NOMA and SDMA in the overloaded network settings averaged over 100 random channel realisations. Blocklength is 250. $K = 4$. $N_t = 2$. (a) is the MMF versus transmit SNR; (b) is the relative gain of RSMA at the transmit SNR of 20 dB.

In Fig. 4.5(a), with underloaded network settings, FBL RSMA has the best performance among the three transmission schemes. The MMF of FBL RSMA with only one splitting user already has a gain of approximately 4.5% and 30% over that of FBL NOMA and FBL SDMA, respectively shown in Fig. 4.5(b). Besides, note that FBL RSMA with more splitting users results in greater MMF. Specifically, increasing the number of splitting users from 1 to 4 brings a fairness rate gain of around 7% and 34% over FBL NOMA and FBL SDMA.

The overloaded case is illustrated in Fig. 4.6. Similarly to the case with $K = 2$, the performance of FBL SDMA deteriorates greatly in Fig. 4.6(a). As shown in Fig. 4.6(b), FBL RSMA has around 100% gain over FBL SDMA when the blocklength is short (e.g. 200). Compared to the underloaded scenario, the gain of FBL RSMA over FBL NOMA and FBL SDMA increases in the overloaded scenario no matter with short or long blocklength.

4.5.2 Link-Level Simulation Results

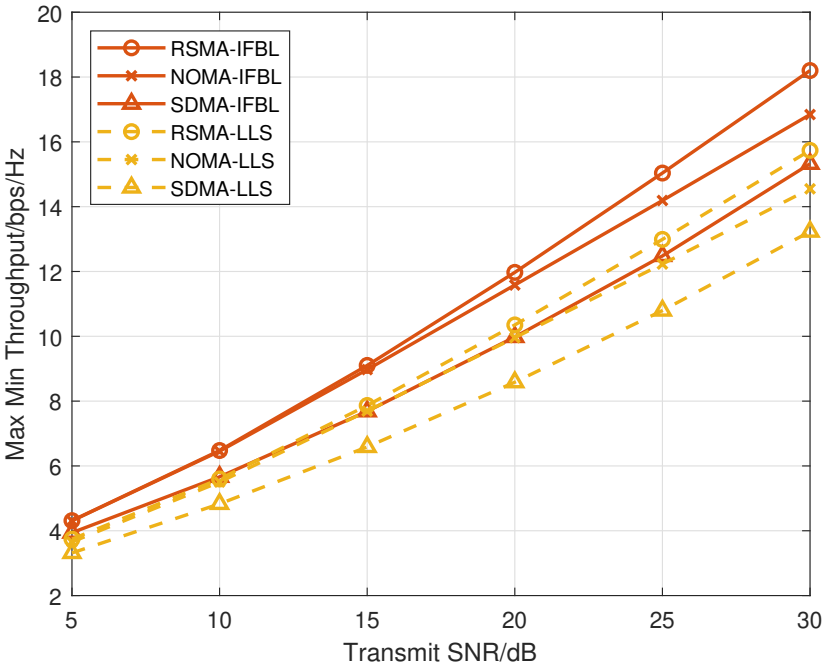
The LLS is performed in this section to analyse the max-min throughput performance of FBL RSMA and compare it with that of FBL NOMA and FBL SDMA, employing the practical transceiver architecture described in Sec. 4.4. An uplink MIMO with 2 users is considered and each has $N_t = 2$ transmit antennas. The number of message stream is $L = \min\{N_t, N_r\}$.

With the assumption of perfect CSIT and CSIR for Modulation and Coding Scheme (MCS) selection, $S^{(l)}$ and $D_k^{(l)}$ are used to denote the number of channel uses in the l^{th} Monte-Carlo realisation and the number of successfully recovered information bits by user-k, respectively. Then, the throughput is calculated as (over 100 channel

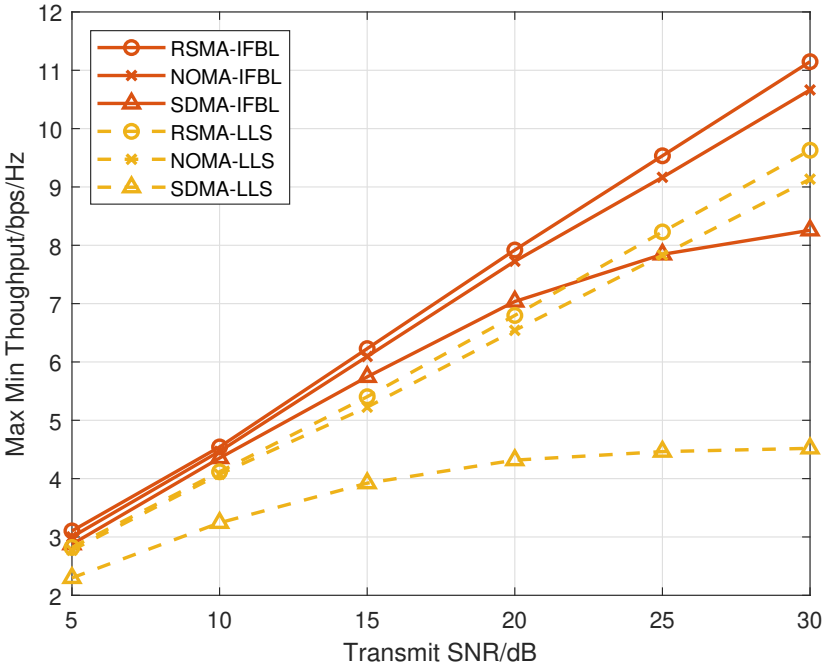
realisations)

$$\text{Max-Min Throughput(bps/Hz)} = \frac{\sum_{l=1}^{100} \min_{k\in\mathcal{K}} D_k^{(l)}}{\sum_{l=1}^{100} S^{(l)}}. \tag{4.31}$$

The modulated blocklength is fixed to $S^l = 256$ throughout the simulations. Fig. 4.7 shows the MMF with IFBL and max-min throughput levels achieved by FBL RSMA, FBL NOMA and FBL SDMA in both overloaded and underloaded scenarios for $N_t = 2$ and $K = 2$. The max-min throughput performance in LLS is consistent with the MMF performance for all three schemes in both underloaded and overloaded regimes that FBL RSMA achieves a significant max-min throughput gain over FBL NOMA and FBL SDMA. However, with the consideration of coding and modulation, the theoretical max-min rate is not achievable, i.e., the throughput of FBL RSMA cannot outperform that of IFBL NOMA as it is the case in Sec. 4.1.



(a) $N_r = 4$



(b) $N_r = 2$

Fig. 4.7: The max-min throughput of RSMA, NOMA and SDMA versus transmit SNR averaged over 100 random channel realisations. (a) is the underloaded scenario; (b) is the overloaded scenario.

4.6 Summary

To conclude, a general K-user RSMA framework in uplink MIMO with perfect CSIT and CSIR is introduced in this chapter, where multiple modulated information streams are transmitted by users compared to only one stream transmitted in the SISO case. The benefits brought by RSMA with FBL constraints are investigated compared to NOMA and SDMA in terms of max-min user rate. Specifically, an AO-based method is proposed to alternatively optimise the precoders at the transmitter side and the combiners at the receiver side, with each solved by SCA. A low-complexity decoding order is also introduced method at the receiver side. Extensive numerical results demonstrate that FBL RSMA can obtain a greater MMF than FBL NOMA and FBL SDMA for a given blocklength. With the increase of blocklength, FBL RSMA can perform better than IFBL NOMA and IFBL SDMA. In multiple-user systems, FBL RSMA with one splitting user can have relatively high gain over FBL NOMA and FBL SDMA. The gain of RSMA becomes significant under the overloaded network settings. With multiple users being split, the gain of FBL RSMA increases. Moreover, the PHY-layer architecture of RSMA in uplink is designed and the performance is analysed in practice. According to the LLS results, it is demonstrated that FBL RSMA in practical settings can still outperform FBL NOMA and SDMA in MIMO.

In summary, RSMA consistently achieves better user fairness than NOMA regardless of FBL or IFBL and RSMA can provide great performance. Increasing the number of splitting users can improve the performance of RSMA at the cost of more sophisticated hardware and processing. In addition, RSMA can achieve the same MMF as NOMA and SDMA but at a shorter blocklength, which achieves significantly lower latency. Therefore, it is concluded that FBL RSMA is a strong and promising PHY-layer approach for networks with multiple antennas.

Chapter 5

RSMA with Mobile Edge Computing

This chapter proposes an RSMA scheme to assist a MEC system where local computation tasks from two users are offloaded to the MEC server facilitated by uplink RSMA for processing. The efficiency of the MEC service is hence primarily influenced by the RSMA-aided task offloading phase and the subsequent task computation phase, where reliable and low-latency communication is required. Therefore, short-packet communication in the FBL regime is taken into consideration. An uplink RSMA-aided MEC framework is proposed and the overall SCP of the service is derived. The offloading factor, transmit power allocation and task-splitting factor are optimised to maximise the SCP with FBL constraints. To deal with the strong coupling between these variables in the SCP expression, the SCA method is applied to decompose the optimisation problem into subproblems. The resultant non-convex subproblems are then tackled by SCA-based algorithm. Numerical results demonstrate that applying uplink RSMA in the MEC system with FBL constraints can not only improve the SCP performance but also provide lower latency in comparison to conventional transmission scheme such as NOMA.

5.1 Introduction

MEC is a promising solution for computation-intensive and latency-critical applications since it allows mobile users to offload their tasks to edge servers for processing, which improves computing performance and decreases network latency compared to remote cloud processing [147, 148]. FBL codes can help with the low-latency requirements in URLLC systems, analysing the performance of MEC in the short-packet communication system is therefore considered to be necessary [6, 7]. Some research has been carried out on MEC in FBL regime [136–138] in terms of EE, reliability, and latency, respectively. However, all those works are based on TDMA or NOMA transmission schemes while uplink RSMA has been demonstrated that it can provide significant gains compared to NOMA in Chapter 3.

This chapter, motivated by exploring the benefits of uplink RSMA in FBL regime in Chapter 3 and to capture the crucial feature of latency-critical requirements of MEC, proposes an uplink RSMA-assisted MEC framework in a two-user scenario and investigates its performance with the constraint of FBL in terms of SCP performance. The main contributions are summarised as follows:

- First, for the *first time*, uplink RSMA with short-packet communications is introduced into MEC system as a powerful MA scheme. The practical study of offloading task data with FBL in this chapter is noteworthy since it is an essential aspect for low-latency applications in 5G and future wireless networks.
- Second, the SCP performance of RSMA-aided MEC under FBL constraints is investigated. This work aims to maximise SCP performance by optimising the entire system strategy with FBL considerations, including the offloading factor, representing the number of tasks to be processed by the MEC server, the task-splitting factor, representing the number of offloading tasks assigned

to each stream in the RSMA scheme and the power allocated to each splitting stream. A closed-form expression of the offloading factor is derived. Due to the coupling among the optimisation variables, AO is applied to decompose the original problem into several sub-problems, each addressed by the proposed SCA-based method. The power allocation and task-splitting factor are designed for the original problem. This is the first work to offer a comprehensive optimisation of systemic parameters for RSMA-aided MEC systems with FBL constraints.

- Through simulations, the performance of RSMA-aided MEC system is analysed. Simulation results show that splitting messages in RSMA helps in MEC deployments because of the flexibility of the decoding order to balance the SINR between each stream. Moreover, RSMA can achieve a higher SCP than NOMA with a shorter blocklength, thereby reducing the latency. However, the performance gain of RSMA over NOMA in the FBL regime saturates as the blocklength becomes sufficiently large. Numerical results demonstrate RSMA-aided MEC achieves more reliable communication in FBL regime with higher SCP compared to NOMA.

5.2 System Model

This section models an RSMA-aided SISO MEC system with two users, indexed by $\mathcal{K} = \{1, 2\}$. Due to the limited local computation capabilities of the two users, this section characterises a MEC scheme that schedules both users to simultaneously offload their data to an MEC server with high computation capability. Facilitated by the MEC server, both users can accomplish their tasks within the time budget by downloading the results from the MEC server. Specifically, it is assumed that each user, with a computational capability of f_{user} Hz/cycle, needs to complete a

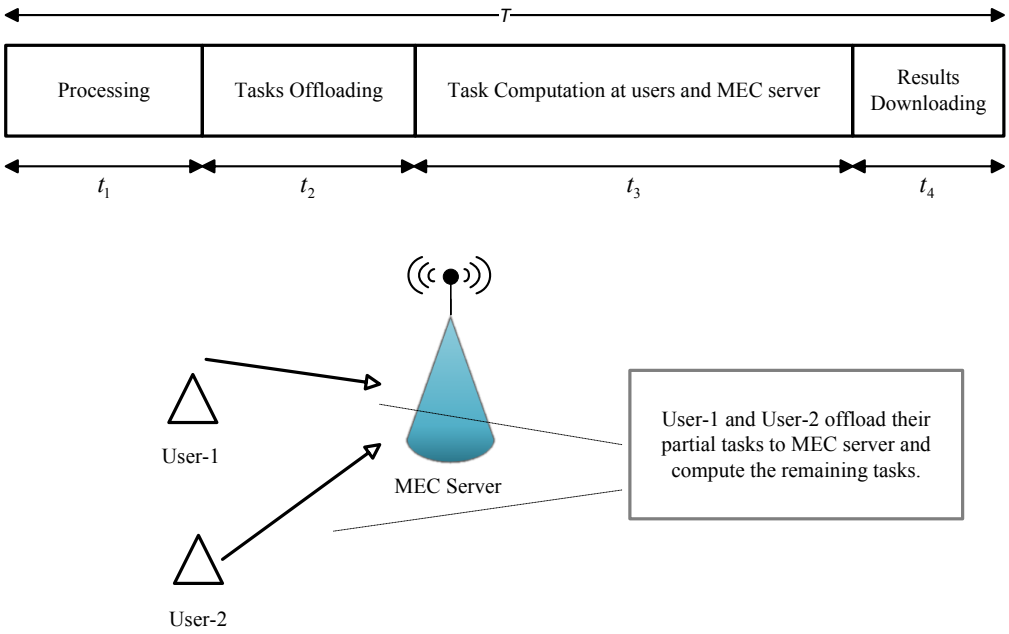


Fig. 5.1: The RSMA-aided MEC system with two users.

computation task of M_k bits within T s. In practice, the scenario often arises that $M_k C_{\text{cpu}}/f_{\text{user}} > T$, where C_{cpu} represents the number of required CPU cycles to compute one-bit task, necessitating the involvement of a central MEC server with a computational capability of $f_{\text{MEC}} = Lf_{\text{user}}$ $(L > 1)$. Assume that the M_k-bits task at each user is bit-wise independent and can be arbitrarily divided into subtasks. Hence, to meet the time budget T, each user-k offloads a part of their task of $\lambda_k M_k$ bits, with λ_k, $0 \leq \lambda_k \leq 1$, being the offloading factor of user-k, to the MEC server with the assistance of uplink RSMA. The remaining tasks of $(1 - \lambda_k)M_k$ bits are executed locally to exploit users' own local computation capability. $\lambda_k = 0$ and $\lambda_k = 1$ represent the fully local computation and fully offloading computation, respectively.

The whole procedure of the RSMA-aided MEC system is shown in Fig. 5.1, which consists of four stages. In the first stage, the offloading, task-splitting factor and power allocation are determined for two users within the processing time t_1 s. The task-splitting factor and power allocation guarantee the error probability performance of uplink RSMA in the offloading stage. Then, in the second stage, users

offload their tasks to the MEC server by uplink RSMA transmission within time duration t_2 s. Thereafter, the third stage performs task execution at both the MEC server and the two local users within time duration t_3 s. It is assumed that the MEC server starts to compute the tasks immediately after the offloading is finished, i.e., there is no queue delay at the MEC server. In the last stage, the computed results of the offloaded tasks are downloaded by users, which consumes time t_4 s. Compared to the offloading time t_2 and execution time t_3, t_1 can be ignored. t_4 can also be ignored as the resultant data often features a very small size compared with the offloading data size in t_2 s [149]. Therefore, the time of offloading computation mainly comes from the time of tasks offloading stage t_2 and the time of tasks execution stage t_3, i.e., in the following of this chapter, it is set that the time constraint to be $t_2 + t_3 \leq T$ without loss of generality.

In the reminder of this section, details of tasks offloading phase (stage 2) in subsection 5.2.1, and the execution parameters in tasks computation phase (stage 3) in subsection 5.2.2 are discussed, respectively. Upon this, the successful offloading probability of stage 2, the successful execution probability of stage 3 and the successful computation probability in Section 5.3 can be formulated.

5.2.1 Tasks Offloading Phase

In this subsection, the details of tasks offloading from two users to the MEC server are presented.

5.2.1.1 RSMA

The two-user uplink SISO RSMA with perfect CSIT and CSIR system is considered. In this system, one user's task is split following the uplink RSMA principle. Assume user-1 split its task W_1 into two parts, $W_{1,i},\ i \in \{1, 2\}$ and the task of user-2 stays

unsplit. These three split tasks $W_{1,1}$, $W_{1,2}$, and W_2 then are encoded independently, resulting into three streams $s_{1,1}$, $s_{1,2}$ and s_2, to be transmitted in total. The received signal at the MEC server is expressed as

$$y = h_1\sqrt{P_{1,1}}s_{1,1} + h_1\sqrt{P_{1,2}}s_{1,2} + h_2\sqrt{P_2}s_2 + n, \tag{5.1}$$

where $P_{1,i}$ is the transmit power of $s_{1,i}, i \in (1,2)$, respectively and P_2 is the transmit power of s_2. They are limited by the power budget that $\sum_{i=1}^{2} P_{1,i} \leq P_t$ and $P_2 \leq P_t$. The scalar $n \sim \mathcal{CN}(0, \sigma_n^2)$ is the complex AWGN at the MEC server.

At the MEC server, SIC is applied to decode three streams and according to Chapter 3 [106], uplink RSMA with the decoding order (i) can provide a much lower error probability compared to NOMA in short-packet communication. Therefore, the decoding order $s_{1,1} \to s_2 \to s_{1,2}$ is adopted in the following demonstration.

Since $s_{1,1}$ is decoded first by treating other signals as noise, the SINR $\gamma_{1,1}$ of the first decoded stream $s_{1,1}$ is

$$\gamma_{1,1} = \frac{P_{1,1}|h_1|^2}{P_{1,2}|h_1|^2 + P_2|h_2|^2 + \sigma_n^2}. \tag{5.2}$$

Assuming $s_{1,1}$ is successfully decoded, its reconstruction $\hat{W}_{1,1}$ is subtracted from the original received signal y to obtain y'. Then, the SINR of the second decoded stream s_2 is

$$\gamma_2 = \frac{P_2|h_2|^2}{P_{1,2}|h_1|^2 + \sigma_n^2}. \tag{5.3}$$

Once $s_{1,2}$ is successfully decoded, its reconstructed task $\hat{W}_{1,1}$ is subtracted from the received signal y'. Thus, the SINR of the last decoded stream $s_{1,2}$ is expressed as

$$\gamma_{1,2} = \frac{P_{1,2}|h_1|^2}{\sigma_n^2}. \tag{5.4}$$

Finally, if $s_{1,2}$ is successfully decoded, the estimated task $\hat{W}_1$ for user-1 is obtained by combining $\hat{W}_{1,1}$ and $\hat{W}_{1,2}$.

The data size of offloading tasks from three transmitted streams is defined as $M_{1,1} = \beta\lambda_1 M_1$, $M_2 = \lambda_2 M_2$ and $M_{1,2} = (1-\beta)\lambda_1 M_1$, where β is the task-splitting factor between splitting streams, satisfying $0 \leq \beta \leq 1$. Then, the total data size of received tasks at the MEC server is given as

$$M_{\text{MEC}} = \lambda_1 M_1 + \lambda_2 M_2. \tag{5.5}$$

Therefore, the Shannon capacity for each stream can be expressed as

$$\begin{aligned} C(\gamma_{1,1}) &= \log_2(1+\gamma_{1,1}), \\ C(\gamma_2) &= \log_2(1+\gamma_2), \\ C(\gamma_{1,2}) &= \log_2(1+\gamma_{1,2}). \end{aligned} \tag{5.6}$$

Due to the consideration of uplink RSMA with FBL, the total transmission time for offloading is

$$t_2 = NT_s, \tag{5.7}$$

where N is the blocklength (in symbol) and T_s is the symbol duration. It is assumed that the blocklength N for each stream is the same.

5.2.1.2 NOMA

NOMA is a particular instance of the system model of Section 5.2.1.1, since $0 \leq P_{1,1} \leq P_1$, NOMA can be regarded as a subset of RSMA. The tasks W_1 and W_2 from user-1 and user-2 are encoded into s_1 and s_2, respectively. The received signal at the MEC server is expressed as $y = h_1\sqrt{P_1}s_1 + h_2\sqrt{P_2}s_2 + n$.

5.2.2 Tasks Computation Phase

In this subsection, the details of the hybrid computations are discussed which are the local computation at the users and the offloading computation at the MEC server, respectively.

For the local computation at user-k, the remaining data size of the tasks is $(1-\lambda_k)M_k$, then the required computation time at user-k is

$$t_3^k = \frac{(1-\lambda_k)M_k C_{\text{cpu}}}{f_{\text{user}}}, \tag{5.8}$$

where C_{cpu} represents the number of CPU cycles required for one-bit task computing. The parameter f_{user} denotes the computation capabilities of users and it is assumed the capacity of each user is the same.

For offloading computation at the MEC server, it is assumed that the computation capability of the MEC server $f_{\text{MEC}} = Lf_{\text{user}}$ with $L > 1$ to indicate the difference between users' and MEC server's computational capabilities. Thus, the required computation time at the MEC server is

$$t_3^{\text{MEC}} = \frac{M_{\text{MEC}} C_{\text{cpu}}}{f_{\text{MEC}}} = \frac{M_{\text{MEC}} C_{\text{cpu}}}{Lf_{\text{user}}}. \tag{5.9}$$

Thus, the time duration t_3 for tasks computation is expressed as

$$\begin{aligned} t_3 &= \max\{t_3^k,\ t_3^{\text{MEC}}\} \\ &= \max\{\frac{(1-\lambda_k)M_k C_{\text{cpu}}}{f_{\text{user}}},\ \frac{M_{\text{MEC}} C_{\text{cpu}}}{Lf_{\text{user}}}\}. \end{aligned} \tag{5.10}$$

5.3 Successful Computation Probability Analysis

In this section, the successful computation probability of RSMA-aided-MEC system with FBL is characterised. The successful computation probability P_s is defined as the probability of the case that all the offloaded tasks and the remaining tasks are computed successfully at the MEC server and the users respectively with the delay budget T. Thus P_s contains two parts which are the successful offloading probability, P_s^o and the successful execution probability, P_s^e, respectively and can be calculated as $P_s = P_s^o P_s^e$.

5.3.1 Successful Offloading Probability P_s^o

The maximal achievable rate expression for FBL with a given error probability ϵ, is given by [61]

$$R \approx \log(1+\gamma) - \sqrt{\frac{V(\gamma)}{N}} Q^{-1}(\epsilon), \tag{5.11}$$

where $V(\gamma) = \left(1 - (1+\gamma)^{-2}\right)$ is the channel dispersion, N is the blocklength, Q is the Q-function and ϵ is the error prorbability. The corresponding coding rate is $R = \frac{M}{N}$. The error probability for a given data size M can be derived based on (5.11) as

$$\epsilon = Q\left(\frac{\log_2(1+\gamma) - \frac{M}{N}}{\sqrt{V(\gamma)/N}}\right) \triangleq Q(f(\gamma, M)), \tag{5.12}$$

where $f(\gamma, M) \triangleq \left(\log_2(1+\gamma) - \frac{M}{N}\right) / \sqrt{V(\gamma)/N}$.

For uplink RSMA, the error probability can be written for each event listed above as: a) $s_{1,1}$ is incorrectly decoded; b) $s_{1,1}$ is correctly decoded but s_2 is incorrectly decoded; c) $s_{1,1}$ and s_2 are both correctly decoded but $s_{1,2}$ is incorrectly decoded.

The error probability of every event can be expressed as

$$\begin{aligned}\epsilon_a &= Q(f(\gamma_a, M_a)),\\ \epsilon_b &= Q(f(\gamma_b, M_b)),\\ \epsilon_c &= Q(f(\gamma_c, M_c)),\end{aligned} \tag{5.13}$$

where $\gamma_a = \gamma_{1,1},\ \gamma_b = \gamma_2$ and $\gamma_c = \gamma_{1,2}$. To simplify the analysis in the following Sec. 5.4, the subscript of M_k and $\beta_k,\ k \in \{1, 2\}$ are changed to align with the subscript notation of three splitting streams. The parameter $M_i,\ i \in \{a, b, c\}$, represents the data size of tasks transmitted by each stream which are $M_a = \beta_a\lambda_1 M_1,\ M_b = \beta_b\lambda_2 M_2$ and $M_c = \beta_c\lambda_1 M_1$, respectively. The parameter $\beta_i,\ i \in \{a, b, c\}$, represents the task-splitting factor between splitting streams where $\beta_a = \beta,\ \beta_b = 1$ and $\beta_c = 1 - \beta$. Then, the error probability of each user can be calculated as

$$\begin{aligned}\epsilon_1 &= \epsilon_a + (1-\epsilon_a)\epsilon_b + (1-\epsilon_a)(1-\epsilon_b)\epsilon_c\\ &\approx \epsilon_a + \epsilon_b + \epsilon_c,\\ \epsilon_2 &= \epsilon_a + (1-\epsilon_a)\epsilon_b \approx \epsilon_a + \epsilon_b.\end{aligned} \tag{5.14}$$

All the multiplied terms, e.g. $\epsilon_a\epsilon_b$ in (5.14) are ignored because the value is negligible. Therefore, the successful offloading probability P_s^o is given by

$$P_s^o = (1-\epsilon_1)(1-\epsilon_2). \tag{5.15}$$

5.3.2 Successful Execution Probability P_s^e

As mentioned in Sec. 5.2.1, the transmission time for offloading is calculated as $t_2 = NT_s$. When $t_3 \le T - t_2$, the successful execution probability P_s^e is 1 or $P_s^e = 0$ if $t_3 \ge T - t_2$ and can be written as $P_s^e = \mathbb{1}(t_3 \le T - t_2)$.

Therefore, the successful computation probability P_s can be written as

$$P_s = \begin{cases} P_s^o, & t_3 \leq T - t_2 \\ 0, & t_3 \geq T - t_2 \end{cases} \tag{5.16}$$

Remark 1: If $\frac{M_k C_{\text{cpu}}}{f_{\text{user}}} \leq T$ that both users can finish their tasks locally within the delay budget so that $P_s = 1$ and offloading is not needed.

5.4 Problem Formulation and Algorithm

In this section, the aim is to maximise the successful computation probability P_s by optimising rate-splitting power allocation for splitting user, $P_{1,1},\ P_{1,2}$, the offloading factor, $\lambda_k, k \in \{1, 2\}$, and task-splitting factor, $\beta_i,\ i \in \{a, b, c\}$, when offloading is performed. Thus, the problem is formulated as

$$\max_{\mathbf{m},\boldsymbol{\lambda},\boldsymbol{\beta}} \quad P_s \tag{5.17a}$$

$$\text{s.t.} \quad P_{1,1} + P_{1,2} \leq P_t, \tag{5.17b}$$

$$P_2 \leq P_t, \tag{5.17c}$$

$$0 \leq \beta_i \leq 1,\ i \in \{a, b, c\} \tag{5.17d}$$

$$t_2 \leq T - t_3, \tag{5.17e}$$

where $\mathbf{m} = [P_{1,1},\ P_{1,2},\ P_2]^T$ represents the transmit power of each stream and $\boldsymbol{\beta} = [\beta_a,\ \beta_b,\ \beta_c]^T$ denotes task-splitting factor between the splitting streams ($\beta_a = 1$).

From (5.15), it is known that

$$\begin{aligned} P_s^o &= (1-\epsilon_1)(1-\epsilon_2) \\ &\approx 1-\epsilon_1-\epsilon_2 \\ &\approx 1-(2\epsilon_a+2\epsilon_b+\epsilon_c) \end{aligned} \tag{5.18}$$

Thus, Problem (5.17) is transformed into

$$\min_{\mathbf{m},\boldsymbol{\lambda},\boldsymbol{\beta}} \quad 2\epsilon_a+2\epsilon_b+\epsilon_c \tag{5.19a}$$

$$\text{s.t.} \quad (5.17\text{b}), (5.17\text{c}), (5.17\text{d}), (5.17\text{e})$$

According to the Chernoff bound of Q-function $\exp(-x^2/2) \geq Q(x), x \geq 0.5$, it is known that

$$\exp\left(\frac{-f^2(\gamma_i, M_i)}{2}\right) \geq Q(f(\gamma_i, M_i)),\ i \in \{a,b,c\}. \tag{5.20}$$

Therefore, Problem (5.19) is transferred into Problem (5.21) to minimise the upper bound of the objective function which is shown as

$$\min_{\mathbf{m},\boldsymbol{\lambda},\boldsymbol{\beta}} \quad 2\exp\left(\frac{-f^2(\gamma_a, M_a)}{2}\right)+2\exp\left(\frac{-f^2(\gamma_b, M_b)}{2}\right)+\exp\left(\frac{-f^2(\gamma_c, M_c)}{2}\right) \tag{5.21a}$$

$$\text{s.t.} \quad (5.17\text{b}), (5.17\text{c}), (5.17\text{d}), (5.17\text{e}).$$

Since Problem (5.17) is not convex and for this kind of multi-variable-optimisation problem, the classical AO algorithms have been widely applied to decompose the original problem into several sub-problems to update each variable in a random order until convergence.

5.4.1 Offloading factor Optimisation

In this subsection, $\boldsymbol{\lambda}$ is optimised under fixed power allocation factor $\mathbf{m}$ and task-splitting factor $\boldsymbol{\beta}$. The subproblem can be written as

$$\min_{\boldsymbol{\lambda}} \quad 2\exp\left(\frac{-f^2(\gamma_a, M_a)}{2}\right) + 2\exp\left(\frac{-f^2(\gamma_b, M_b)}{2}\right) + \exp\left(\frac{-f^2(\gamma_c, M_c)}{2}\right) \tag{5.22a}$$

$$\text{s.t.} \quad t_3 \leq T - t_2. \tag{5.22b}$$

According to (5.7) and (5.10), constraint (5.22b) can be rewritten as

$$\frac{(1-\lambda_k)M_k C_{\text{cpu}}}{f_{\text{user}}} \leq T - NT_s, \ k \in \{1,2\} \tag{5.23a}$$

$$\frac{\sum_{k=1}^{2}\lambda_k M_k C_{\text{cpu}}}{L f_{\text{user}}} \leq T - NT_s, \ k \in \{1,2\}. \tag{5.23b}$$

Therefore, Problem (5.22) can be transformed into

$$\min_{\boldsymbol{\lambda}} \quad 2\exp\left(\frac{-f^2(\gamma_a, M_a)}{2}\right) + 2\exp\left(\frac{-f^2(\gamma_b, M_b)}{2}\right) + \exp\left(\frac{-f^2(\gamma_c, M_c)}{2}\right) \tag{5.24a}$$

$$(5.23a), (5.23b).$$

Lemma 1: With fixed power allocation factor $\mathbf{m}$ and task-splitting factor $\boldsymbol{\beta}$, the closed-form for λ_k is given by

$$\lambda_k^\star = \max\{0, \ 1 - \frac{(T - NT_s)f_{\text{user}}}{M_k C_{\text{cpu}}}\}, \ k \in \{1,2\}. \tag{5.25}$$

Proof: For arbitrary $\mathbf{m}$ and $\boldsymbol{\beta}$, the objective function $2\exp\left(\frac{-f^2(\gamma_a,M_a)}{2}\right)$ $+2\exp\left(\frac{-f^2(\gamma_b,M_b)}{2}\right)+\exp\left(\frac{-f^2(\gamma_c,M_c)}{2}\right)$ is clearly monotonically increasing with respect to (w.r.t) $\exp\left(\frac{-f^2(\gamma_i,M_i)}{2}\right)$, $i \in \{a,b,c\}$. Besides, each $\exp\left(\frac{-f^2(\gamma_i,M_i)}{2}\right)$ in the objective function is monotonically decreasing w.r.t λ_k. Therefore, the objective function in

Problem (5.22) is monotonically increasing w.r.t λ_k. From (5.22b), it is able to obtain the lower bound of λ_k

$$1 - \frac{(T - NT_s)f_{\text{user}}}{M_k C_{\text{cpu}}} \leq \lambda_k,\ k \in \mathcal{K}. \tag{5.26}$$

If $1 - \frac{(T-NT_s)f_{\text{user}}}{M_k C_{\text{CPU}}} \leq 0$, then $T - NT_s \geq \frac{M_k C_{\text{cpu}}}{f_{\text{user}}}$. It means each user can finish their own tasks within the time budget and it is not necessary to offload tasks to MEC server so that $\lambda_k^* = 0$. As the objective function is monotonically increasing, the optimal value of λ_k should be $\lambda_k^\star = \max\{0,\ 1 - \frac{(T-NT_s)f_{\text{user}}}{M_k C_{\text{cpu}}}\},\ k \in \{1, 2\}$.

5.4.2 Transmit Power Optimisation

For given feasible $\boldsymbol{\beta}$ and offloading factor $\boldsymbol{\lambda}$, by introducing the slack variables $\mathbf{t} = [t_a,\ t_b,\ t_c]^T$ and $\boldsymbol{\rho} = [\rho_a,\ \rho_b,\ \rho_c]^T$, the subproblem is formulated as

$$\min_{\mathbf{m},\boldsymbol{\rho}} \quad 2\exp(-t_a) + 2\exp(-t_b) + \exp(-t_c) \tag{5.27a}$$

$$\text{s.t.} \quad t_i \leq \frac{N[\log(1+\rho_i) - \frac{M_i}{N}]^2}{2[1-(1+\rho_i)^{-2}]}, i \in \{a, b, c\} \tag{5.27b}$$

$$\frac{P_{1,1}|h_1|^2}{P_{1,2}|h_1|^2 + P_2|h_2|^2 + \sigma_n^2} \geq \rho_a, \tag{5.27c}$$

$$\frac{P_2|h_2|^2}{P_{1,2}|h_1|^2 + \sigma_n^2} \geq \rho_b, \tag{5.27d}$$

$$\frac{P_{1,2}|h_1|^2}{\sigma_n^2} \geq \rho_c, \tag{5.27e}$$

$$(5.17\text{b}),\ (5.17\text{c}).$$

Constraint (5.27b) is not convex, it is firstly rewritten it into

$$\frac{t_i}{N}(1-(1+\rho_i)^{-2}) \leq \frac{1}{2}\left(\log(1+\rho_i) - \frac{M_i}{N}\right)^2. \tag{5.28}$$

Since (5.30) is not convex, the Right-Hand Side (RHS) is approximated by its fist-order Taylor expression which is given by

$$\frac{1}{2}\left(\log(1+\rho_i)-\frac{M_i}{N}\right)^2 \geq a_i^{[n]}\log(1+\rho_i)+b_i^{[n]}, \tag{5.29}$$

where $a_i^{[n]}=\log(1+\rho_i^{[n]})-\frac{M_i}{N}$ and $b_i^{[n]}=\frac{1}{2}(a_i^{[n]})^2-a_i^{[n]}\log(1+\rho_i^{[n]})$. An additional auxiliary variable is introduced, $\mathbf{t}_1=[t_{1,a},\ t_{1,b},\ t_{1,c}]^T$, to approximate the LHS of constraint (5.26) and the following constraints for the introduced variable is added:

$$t_{1,i}\leq t_i(1+\rho_i)^{-2}. \tag{5.30}$$

Then (**??**) is transformed into

$$\log(t_{1,i})+2\log(1+\rho_i)\leq\log(t_i). \tag{5.31}$$

Since (5.31) is still not convex, $\log(t_{1,i})$ and $2\log(1+\rho_i)$ are approximated by their first-order approximation around the point $\left(t_{1,i}^{[n]}\right)$ and $\left(\rho_i^{[n]}\right)$, which are given by

$$\log(t_{1,i})\leq\log(t_1^{[n]})+\frac{1}{t_{1,i}^{[n]}}(t_{1,i}-t_{1,i}^{[n]}) \tag{5.32a}$$

$$2\log(1+\rho_i)\leq 2\log(1+\rho_i^{[n]})+\frac{2}{1+\rho_i^{[n]}}(\rho_i-\rho_i^{[n]}). \tag{5.32b}$$

Then the constraint (5.30) can be rewritten into

$$\frac{t_i}{N}-\frac{t_{1,i}}{N}\leq a_i^{[n]}\log(1+\rho_i)+b_i^{[n]} \tag{5.33a}$$

$$\log(t_1^{[n]})+\frac{1}{t_{1,i}^{[n]}}(t_{1,i}-t_{1,i}^{[n]})+2\log(1+\rho_i^{[n]})+\frac{2}{1+\rho_i^{[n]}}(\rho_i-\rho_i^{[n]})\leq\log(t_i). \tag{5.33b}$$

Additionally, constraints (5.27c) and (5.27d) are not convex. Then, they are equiv-

alently written into Difference-of Convex forms, which are given by

$$P_{1,2}|h_1|^2 + P_2|h_2|^2 + \sigma_n^2 - \frac{P_{1,1}|h_1|^2}{\rho_a} \leq 0, \tag{5.34a}$$

$$P_{1,2}|h_1|^2 + \sigma_n^2 - \frac{P_2|h_2|^2}{\rho_b} \leq 0, \tag{5.34b}$$

and by rewriting the concave part with the first-order Taylor approximations, the constraints (5.34a) and (5.34b) are respectively approximated around the point $\left(\mathbf{m}^{[n]}, \boldsymbol{\rho}^{[n]}\right)$ at iteration n by

$$P_{1,2}|h_1|^2 + P_2|h_2|^2 + \sigma_n^2 - \frac{P_{1,1}|h_1|^2}{\rho_a^{[n]}} + (\rho_a - \rho_a^{[n]})\frac{P_{1,1}^{[n]}|h_1|^2}{(\rho_a^{[n]})^2} \leq 0 \tag{5.35a}$$

$$P_{1,2}|h_1|^2 + \sigma_n^2 - \frac{P_2|h_2|^2}{\rho_b^{[n]}} + (\rho_b - \rho_b^{[n]})\frac{P_2^{[n]}|h_2|^2}{(\rho_b^{[n]})^2} \leq 0. \tag{5.35b}$$

The non-convex subproblem (5.27) is transformed into a convex Problem (5.36) based on the aforementioned approximation techniques, which can then be solved via the SCA approach. By approximating a series of convex sub-problems, SCA solves the original problem. At iteration n, the following problem is solved using the best solution $\left(\mathbf{m}^{[n-1]}, \boldsymbol{\rho}^{[n-1]}\right)$ from the preceding iteration $n-1$:

$$\min_{\mathbf{m},\boldsymbol{\rho}} \quad 2\exp\left(-t_a\right) + 2\exp\left(-t_b\right) + \exp\left(-t_c\right) \tag{5.36a}$$

$$\text{s.t.} \quad (5.33\text{a}), (5.33\text{b}), (5.35\text{a}), (5.35\text{b})(5.27\text{e})(5.17\text{b}), (5.17\text{c}).$$

5.4.3 Task-Splitting Factor Optimisation

When transmit power $\mathbf{m}$ and the offloading factor $\boldsymbol{\lambda}$ are all fixed, the subproblem with respect to $\boldsymbol{\beta}$ is given by

$$\max_{\boldsymbol{\beta}} \quad 2\exp(-t_a) + 2\exp(-t_b) + \exp(-t_c) \tag{5.37a}$$

$$\text{s.t.} \quad t_i \leq \frac{N[\log(1+\rho_i) - \frac{\beta_i \lambda_k M_k}{N}]^2}{2[1-(1+\rho_i)^{-2}]},\ i \in \{a,b,c\},\ k \in \{1,2\} \tag{5.37b}$$

(5.17d).

Remark 3: Notice that in (5.37b), the subscription $i = \{a, c\}$ corresponds to $k = 1$ and $i = b$ corresponds to $k = 2$ based on the definition of M_i in Sec. 5.3.1.

The constraint (5.37b) is rewritten as

$$t_i \leq N c_i (d_i + e_i \beta_i)^2, \tag{5.38}$$

where $c_i \triangleq \frac{1}{2[1-(1+\rho_i)^{-2}]}$, $e_i \triangleq -\frac{\lambda_i M_k}{N}$ and $d_i \triangleq \log(1+\rho_i)$.

Similar to Sec. 5.4.1, constraint (5.38) is not convex, $(e_i\beta_i)^2$ is approximated by its first-order Taylor approximation around the point $\left(\boldsymbol{\beta}^{[n]}\right)$ at iteration n which is given by

$$(e_i\beta_i)^2 \geq (e_i\beta_i^{[n]})^2 + 2(\beta_i - \beta_i^{[n]})(e_i^2\beta_i^{[n]}) \triangleq \Phi(\beta_i^{[n]}). \tag{5.39}$$

Thus, constraint (5.37b) is rewritten into

$$t_i \leq N c_i (d_i^2 + 2 d_i e_i \beta_i + \Phi(\beta_i^{[n]})). \tag{5.40}$$

Therefore, subproblem (5.37) is transformed into

$$\min_{\boldsymbol{\beta}} \quad 2\exp\left(-t_a\right) + 2\exp\left(-t_b\right) + \exp\left(-t_c\right) \tag{5.41a}$$

$$\text{s.t.} \quad (5.40), (5.17\text{d}).$$

Now, it is easy to verify that all these sub-problems, Problem (5.36) and Problem (5.41) are convex problems, which can be efficiently solved by standard convex problem solver such as CVX [143].

According to *Lemma 1* and the above two subproblems, AO is applied to solve Problem (5.21) by utilising the SCA-based algorithm. Specifically, the offloading factor $\boldsymbol{\lambda}$ is calculated by (5.25) and transmit power $\mathbf{m}$ and the task-splitting factor $\boldsymbol{\beta}$ are alternately optimised by solving Problem (5.36) and Problem (5.41), respectively. The details of the proposed algorithm are summarised in Algorithm 4 where τ is the tolerance. Define $\epsilon = 2\exp\left(-t_a\right) + 2\exp\left(-t_b\right) + \exp\left(-t_c\right)$ for convenience and $\epsilon^{[n]} = 2\exp\left(-t_a^{[n]}\right) + 2\exp\left(-t_b^{[n]}\right) + \exp\left(-t_c^{[n]}\right)$.

At each iteration of Algorithm 4, the transmit power and task-splitting factor are updated by solving SOCP problems. Each SOCP is solved by using interior-point method with the computational complexity of $\mathcal{O}\left([X]^{3.5}\right)$, where X is the total number of variables in the corresponding SOCP problem. Although ad-

Algorithm 4: Proposed SCA-based AO algorithm for solving Problem (5.21)

Initialise: $n \leftarrow 0,\ \epsilon^{[n]} \leftarrow 0$, and feasible $\mathbf{m}^{[n]},\ \boldsymbol{\beta}^{[n]}$;
Calculate optimal $\lambda_k^*, k \in \{1, 2\}$ by *Lemma 1*.

repeat
  $n \leftarrow n + 1$;
  Find optimal $\mathbf{m}^{[n]}$ by solving Problem (5.36) with given $\boldsymbol{\beta}^{[n-1]}$;
  Find optimal $\boldsymbol{\beta}^{[n]}$ by solving Problem (5.41) with calculated $\mathbf{m}^{[n]}$;
  Update $\epsilon^{[n]} \leftarrow \epsilon^*,\ \mathbf{m}^{[n]} \leftarrow \mathbf{m}^*,\ \boldsymbol{\beta}^{[n]} \leftarrow \boldsymbol{\beta}^*$;
until $|\epsilon^{[n]} - \epsilon^{[n-1]}| \leq \tau$;

ditional variables are introduced for approximating convex problems, the main complexity still comes from the transmit power and task allocation factor design. With given task allocation factor, the number of variables of Problem (5.41) is given by $X_{\text{transmit power}} = (K+1)$. Similarly, the number of variables of Problem (5.41) is given by $X_{\text{task-splitting factor}} = (K+1)$. The total number of iterations required for the convergence is $\mathcal{O}\left(\log\left(\tau^{-1}\right)\right)$, where τ is the convergence tolerance of Algorithm 4. Therefore, the total computation complexity of Algorithm 4 is $\mathcal{O}\left((K+1)^{3.5}\log\left(\tau^{-1}\right)\right)$.

5.5 Numerical Results and Discussion

The performance of a FBL RSMA-aided MEC system with two users is evaluated and compared with conventional transmission schemes in this section. The SCP performance of RSMA-aided MEC is illustrated. Here, the performance of RSMA is compared to NOMA.

- NOMA: This is a special case of RSMA. As specified in Sec. 3.2.2 of Chapter 3, none of the users split their messages in NOMA. The BS sequentially decodes the user messages based on SIC.

Simulations for the Rayleigh Fading channel with 100 channel realisations are performed. The time budget, T, is 10 ms and the time length of blocklength (in symbol), T_s, is 0.0025 ms. The CPU computation is $C_{\text{cpu}} = 1000$ cycles/bits, $f_{\text{user}} = 0.5$ GHz and L is assumed to be 5. Without loss of generality, it is assumed that the noise variance is 1, $\sigma_n^2 = 1$. The tolerance of the algorithm is set to be $\tau = 10^{-3}$. Here, the decoding orders of $s_{1,1} \to s_2 \to s_{1,2}$ of RSMA and $s_1 \to s_2$ of NOMA are chosen, respectively.

5.5.1 The SCP versus Task Size

The trend of SCP as the size of tasks increases is shown in Fig. 5.2. Results of four different blocklengths, $N = 250, 500, 750$ and 1000, are compared. In this simulation, the task size of user-1, M_1, is from 5k bits to 10k bits and the task size of user-2 is $M_2 = 5.5$k bits. Solid and dash lines represent the transmit SNR of 10 dB and 15 dB, respectively. Recall that when $\frac{M_k C_{\text{cpu}}}{f_{\text{user}}} \leq T,\ k \in 1, 2$, user-$k$ can complete its tasks locally within the time budget, making offloading unnecessary. For user-1, with $M_1 = 5$k bits, the local computation time is $\frac{M_1 C_{\text{cpu}}}{f_{\text{user}}} = 10$ ms $\leq T$, so offloading is not required. In contrast, user-2 has a local computation time of $\frac{M_2 C_{\text{cpu}}}{f_{\text{user}}} = 11$ ms $> T$, necessitating offloading to meet the time budget. Consequently, transmission errors occur only during user-2's offloading process when $M_1 = 5$k bits and $M_2 = 5.5$k bits. As M_1 increases, resulting in longer local computation times for user-1, offloading to the MEC server becomes necessary for user-1. The insights from Fig. 5.2 are summarised as follows.

- Fig. 5.2 firstly illustrates the enhanced SCP performance as SNR increases, comparing the solid line (SNR=10 dB) and the dashed line (SNR=15 dB). For example, in Fig. 5.2 (a), with a short blocklength of 250, the SCP of both RSMA and NOMA drops to 0 when the task size reaches and goes beyond 5.5k bits given a transmit SNR of 10 dB. This is because the short blocklength results in a significantly high coding rate in FBL, which, however, can be mitigated by an increased SNR. As a verification, in Fig. 5.2 (a), when the transmit SNR increases to 15 dB, the SCP is approximately 0.2 (not zero as a comparison) for $M_1 = 5.5$k bits and drops to 0 when $M_1 = 6$k bits.

- Secondly, Fig. 5.2 demonstrates the superiority of RSMA over NOMA in the MEC scheme. For example, Fig. 5.2 (b), with a task size of 6k bits and a transmit SNR of 15 dB, RSMA achieves a significantly higher SCP of

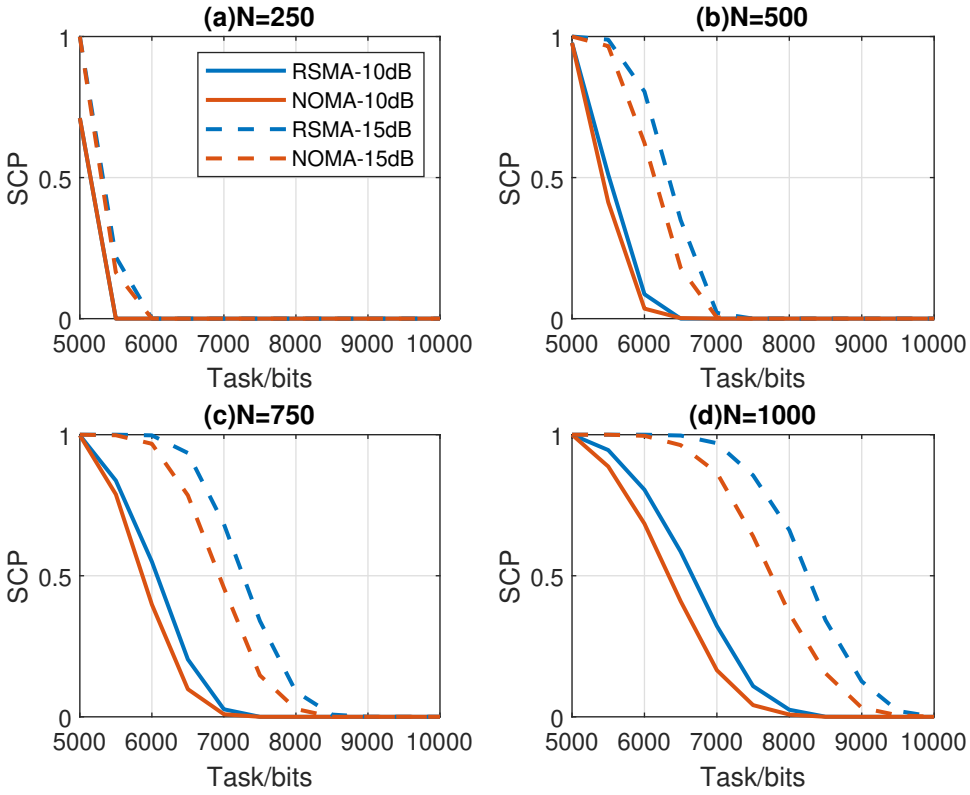


Fig. 5.2: The successful computation probability performance of RSMA and NOMA versus the size of tasks with two different transmit SNR averaged over 100 random channel realisations. M_2 = 5.5k bits. (a) N=250; (b) N=500; (c) N=750; (c) N=1000.

approximately 0.8 with a blocklength of 500, as shown in Fig. 5.2, compared to NOMA, which achieves an SCP of about 0.6. At a SCP of 0.5 and a blocklength of 750, the maximum task sizes for RSMA are approximately 6.1k bits and 7.4k bits, compared to 5.8k bits and 7k bits for NOMA, at transmit SNRs of 10 dB and 15 dB, respectively, as shown in Fig. 5.2 (c).

The superiority of RSMA over NOMA comes from the higher flexibility in decoding order and in allocating power to each stream. The streams of the second decoded user would have a considerable impact on the first decoded user's SINR since the decoding order strictly limits NOMA. However, the decoding order can be more flexible to balance the SINR while decoding each stream with the help of RSMA by splitting one user, for instance, the decoding

order for RSMA is $s_{1,1} \rightarrow s_2 \rightarrow s_{1,2}$. In this approach, the task M_1 for user-1 is split into two parts which are transmitted by $s_{1,1}$ and $s_{1,2}$. The majority of M_1, along with a significant portion of user-1's transmit power, is allocated to $s_{1,2}$, which is decoded interference-free after s_2. This strategy significantly reduces the decoding burden on $s_{1,1}$, enabling more efficient interference management compared to NOMA, where the entire task M_1 must contend with interference. By leveraging task splitting and optimising the power allocation, RSMA achieves better resource utilisation and improves the overall system performance.

- Finally, Fig. 5.2 (a) - Fig. 5.2 (d) demonstrates an enlarged performance gain of RSMA over NOMA with the increase of blocklength. Especially in Fig. 5.2 (d), for RSMA, the maximum task size that user-1 can handle is 6.7k bits at a transmit SNR of 10 dB and 8.2k bits at 15 dB. In comparison, for NOMA, the maximum task size is 6.5k bits at 10 dB and 7.8k bits at 15 dB.

  Additionally, the offloading time, $t_2 = NT_s$, increases with the blocklength, reducing the available computation time, t_3. As a result, the user must offload a larger portion of tasks to the MEC server, leading to a higher coding rate. In this scenario, RSMA proves advantageous by splitting a user's tasks, effectively balancing the SINR of each stream and reducing the offloading error probability.

5.5.2 The SCP versus Blocklength

Fig. 5.3 illustrates the relationship between SCP and blocklength. The comparison is made for two task sizes, $M_k = 6$k bits, $M_k = 7$k bits and $M_k = 8$k bits, where $k \in \{1, 2\}$. The transmit SNR is fixed at 15 dB for this simulation. RSMA requires a shorter blocklength than NOMA to achieve the same SCP. For example, when

the task size for both users is 6k bits, RSMA achieves an SCP of 0.5 with 50 fewer blocklengths (in symbol) compared to NOMA as indicated by the two solid lines. This demonstrates that RSMA can effectively reduce latency by requiring less blocklength. As the task size increases, the SCP performance gain of RSMA over NOMA becomes more significant, as illustrated by the other two groups of lines. For larger task sizes, the coding rate increases proportionally, placing a heavy burden on decoding s_1 at the MEC server, as its SINR is heavily influenced by s_2 for NOMA. To decode s_2 at a high coding rate, more power is allocated to s_2, which results in an increasing interference experienced by s_1.

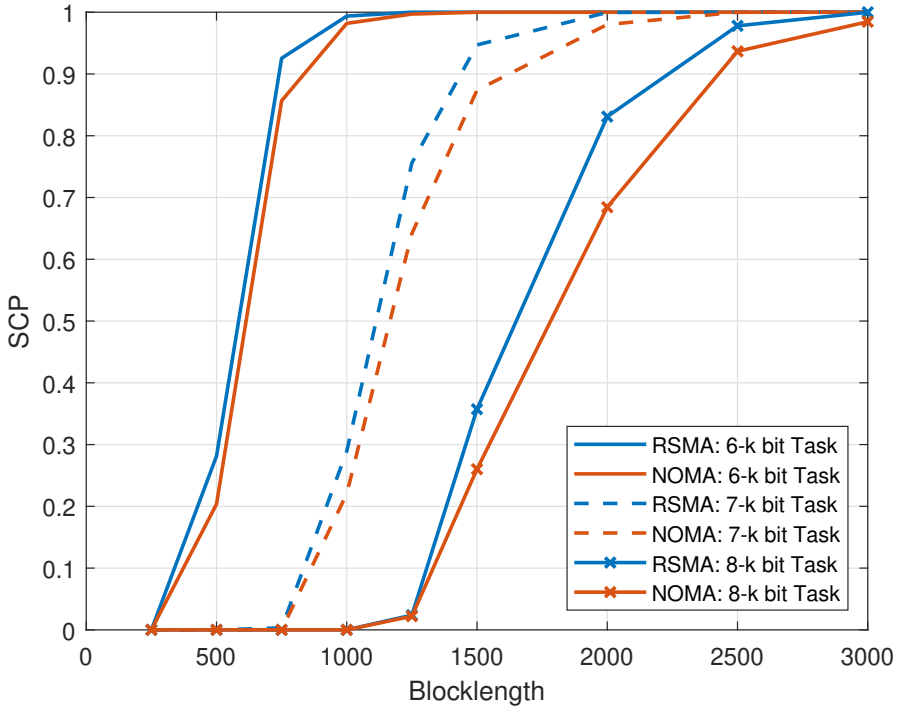


Fig. 5.3: The successful computation probability performance of RSMA and NOMA versus blocklength with two different task sizes averaged over 100 random channel realisations. Transmit SNR is 15 dB.

5.5.3 The SCP versus Transmit SNR

Fig. 5.4 shows the trend of SCP as the transmit SNR increases with more comprehensive simulations. The comparison includes results for five specific blocklengths: $N = 500$, 1000, 1500, 2000 and 3000. In this simulation, both user-1 (M_1) and user-2 (M_2) have task sizes of 7k bits. The blue lines represent RSMA, while the red lines correspond to NOMA.

In Fig. 5.4, with blocklength of 500, neither RSMA and NOMA can successfully complete own tasks within the time budget (SCP is 0) since the coding rate is too high. As the blocklength increases, the SCP performance improves with the SCP of NOMA remains consistently lower than RSMA. The gain of RSMA over NOMA does not continue to increase with blocklength, instead, it saturates at the blocklength of approximately 1000, revealing the performance limit of RSMA.

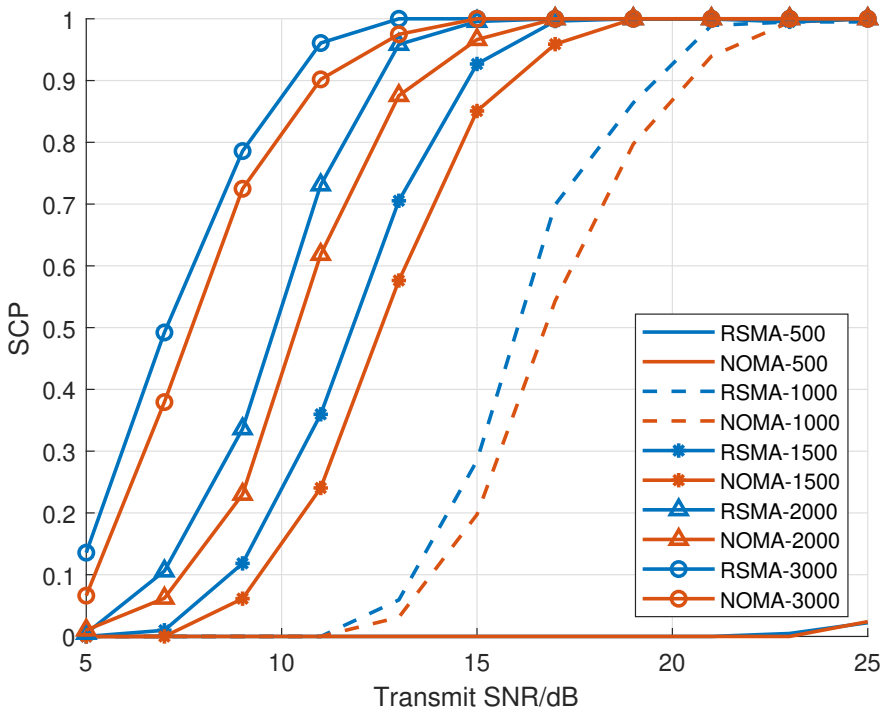


Fig. 5.4: The successful computation probability performance of RSMA and NOMA versus transmit SNR with four different finite blocklength averaged over 100 random channel realisations. $M_1 = 7$k bits. $M_2 = 7$k bits.

These numerical results demonstrates that RSMA with a decoding order of $s_{1,1} \rightarrow s_2 \rightarrow s_{1,2}$ achieves a higher SCP compared to NOMA with a decoding order of $s_1 \rightarrow s_2$ in FBL regime. The superior performance of RSMA can be attributed to its ability to allocate transmit power and offloading task size between split streams, balancing the SINR and enhancing overall SCP performance.

5.6 Summary

This chapter introduces an RSMA-aided MEC system operating in the FBL regime, where SCP serves as a critical metric. The study focuses on a two-user RSMA-aided MEC system, aiming to optimise power allocation, task-splitting factor, and offloading factors to maximise SCP. Emphasising the scenario where only one user's task is split, simulations demonstrate that the RSMA-aided MEC system achieves significantly higher SCP compared to traditional schemes such as NOMA with FBL constraints. Furthermore, RSMA enables reduced latency by achieving comparable SCP to NOMA with shorter blocklengths. These findings shows RSMA-assisted MEC as a promising approach for reliable, low-latency wireless communication systems in the future.

In conclusion, RSMA continuously produces a higher SCP than NOMA in FBL regime, allowing RSMA to provide a more dependable performance. Besides, RSMA achieves much lower latency while achieving the same SCP as NOMA with a shorter blocklength. Since this work is based on a two-user SISO system, the general K-user MIMO system can be studied in the future.

Chapter 6

Conclusions

6.1 Summary of Thesis Achievements

In this thesis, the investigation of the application of uplink RSMA and RSMA-aided MEC system in the FBL regime are studied, addressing several optimisation problems and algorithms across various scenarios. These include a two-user uplink SISO channel system with/without time-sharing, and a K-user uplink MIMO channel system with optimised precoders and combiner followed by the PHY-layer simulations and an RSMA-aided MEC system. Extensive simulation results and analysis are presented to evaluate the performance of the proposed schemes.

In Chapter 3, RSMA in uplink communications is introduced into the FBL regime, catering to low latency requirements. A two-user uplink SISO system is investigated, aiming to optimise power and rate allocation factors for maximising sum-throughput and rate region with/without time-sharing. At the same time, the impact of blocklength and transmission rate on the throughput and error probability performance of RSMA with the message of only one user being split is considered. Through simulations, the superiority of uplink SISO RSMA in the FBL regime is

demonstrated, showing higher sum-throughput at the same transmission rate and a larger rate region under the same error probability constraints compared to conventional schemes, i.e. NOMA. Moreover, RSMA can achieve the same throughput as NOMA with a shorter blocklength, thereby reducing the latency. The results underscore the potential of uplink RSMA as a promising multiple access strategy for future wireless communications characterised by reliability and low latency.

In Chapter 4, the performance of RSMA with multiple transmit and receive antennas under FBL constraints in short-packet uplink communications is investigated. Considering a K-user uplink system, an uplink MIMO RSMA framework is proposed and both precoders and combiners with the MMF metric are optimised. Given the strong coupling between precoders and combiners, AO is employed to decompose the optimisation problem into two subproblems. To tackle these subproblems, an approach based on SCA is proposed. A low-complexity method for decoding order at the receiver is also introduced. Extensive numerical simulations show that under finite FBL constraints, RSMA achieves higher MMF compared to NOMA and SDMA at the same blocklength. Moreover, as the blocklength increases, FBL RSMA outperforms IFBL NOMA and IFBL SDMA. In multi-user scenarios, FBL RSMA with one splitting user exhibits considerable gains over FBL NOMA and FBL SDMA. These gains become more pronounced in overloaded network conditions. Additionally, the benefits of splitting multiple users are also demonstrated. Furthermore, the design and evaluation of the practical PHY-layer architecture of RSMA for uplink scenarios are done. The LLS results demonstrate that in practical settings, FBL RSMA consistently outperforms FBL NOMA and SDMA in MIMO channels, leading us to conclude that uplink MIMO RSMA is superior to existing transmission schemes and has great potential to improve the system performance of future communication networks.

In Chapter 5, RSMA with FBL constraints is introduced into the MEC system,

catering to high SCP and low latency requirements. A two-user RSMA-aided MEC system is investigated, aiming to maximise SCP by optimising power allocation, task-splitting factor, and offloading factor with the task from only one user being split based on the conclusion of Chapter 3. Through simulations, the superiority of RSMA-aided MEC system in the FBL regime is demonstrated, showing higher SCP compared to conventional schemes, i.e. NOMA. Moreover, RSMA can achieve the same SCP as NOMA with a shorter blocklength, thereby reducing the latency. The results prove the potential of RSMA-aided MEC as a promising multiple access strategy and system for future wireless communications characterised by reliability and low latency.

6.2 Future Work

In conclusion of this thesis, some potential future research directions are listed as follows:

1. **RSMA in FBL regime with imperfect CSIT:** In Chapter 3, 4, and 5, the assumption of perfect CSIT and CSIR has been made to shed light on how RSMA performs with FBL. Unfortunately, in practice, CSIT and CSIR are always imperfect due to pilot reuse, channel estimation errors, pilot contamination, limited and quantised feedback accuracy, delay/latency, mobility (due to ever-increasing speeds of vehicle/trains/satellite/flying objects and emerging applications as Vehicle-to-Everything), Radio Frequency (RF) impairments (e.g., phase noise), inaccurate calibrations of RF chains, subband level estimation [31]. Although extensive works are studying downlink RSMA performance with infinite blocklength and imperfect CSIT and the superiority of RSMA over conventional strategies has been shown in terms of CSIT inaccuracy and DoF [35, 150–152], there is no work focusing on uplink RSMA performance with IFBL and imperfect CSIT. Moreover, the performance of uplink FBL RSMA with imperfect CSIT is an open problem worth future investigation.

2. **RSMA for mMTC in FBL regime:** In addition to FBL code for URLLC addressed in Chapter 3, 4 and 5 of this thesis, there is another application which is also important to 5G and beyond communication, namely mMTC. Besides, IoT is essential for 5G and beyond as it provides advanced solutions for smart cities via a massive number of IoT devices. mMTC, which is tailored for massive IoT applications, has been included in 3GPP Release 13 [153] as an instrumental usage scenario for ultimately realising massive connectivity among IoT devices with low complexity and low power. Although

there is one work investigating uplink RSMA with enhanced Mobile Broadband (eMBB)-URLLC-mMTC [57], the performance of uplink FBL RSMA with hybrid requirements is an open problem worth future investigation.

3. **Physical layer security of RSMA:** Privacy/security is one of the most challenging and critical issues in wireless communications due to the broadcast nature of wireless channels. Reference [154] was the first work showing the explicit gain regarding max-min secrecy rate gain of RSMA over SDMA and NOMA. [155] has demonstrated cooperative RS yields a secrecy sum-rate gain over cooperative NOMA and SDMA. Additionally, in Time Division Depulx (TDD) massive MIMO, pilot-contamination attackers may have an impact on the uplink channel estimation by actively transmitting the same pilot sequence as the legitimate users in an effort to diminish the estimated CSIT of the legitimate users. Motivated by RSMA's enhanced robustness to pilot contamination [156], it is worth looking into how well RSMA performs in terms of secrecy rate in TDD massive MIMO to see if it can combat pilot contamination attacks.

References

[1] Huawei, "5G ToB service experience standard white paper," https://carrier.huawei.com/~/media/cnbgv2/download/products/servies/5g-b2b-service-experience-standard-white-paper-en1.pdf, accessed at 19/03/2024.

[2] 3GPP TR 38.913 V17.0.0, "3rd Generation Partnership Project;Technical Specification Group Radio Access Network;Study on Scenarios and Requirements for Next Generation Access Technologies;(Release 17)," 3GPP, techreport, Mar. 2022.

[3] A. Manikas, "Advanced communication theory," accessed at 01/09/2024, [Online], Available: https://skynet.ee.ic.ac.uk/notes/ACT_5_Extended%20Array_Architectures_optimised.pdf.

[4] M. I. Y. Zahraa Qasem Jaber, "Design and implementation of real time face recognition system (rtfrs)," *International Journal of Computer Applications*, vol. 94, no. 12, pp. 15–22, May 2014. [Online]. Available: https://ijcaonline.org/archives/volume94/number12/16395-6014/

[5] T. Verbelen, P. Simoens, F. De Turck, and B. Dhoedt, "Leveraging cloudlets for immersive collaborative applications," *IEEE Pervasive Computing*, vol. 12, no. 4, pp. 30–38, 2013.

[6] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *Ieee Access*, vol. 6, pp. 55 765–55 779, 2018.

[7] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1711–1726, 2016.

[8] Y. Mao, O. Dizdar, B. Clerckx, R. Schober, P. Popovski, and H. V. Poor, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2073–2126, 2022.

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

[10] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *2013 IEEE 24th annual international symposium on personal, indoor, and mobile radio communications (PIMRC)*. IEEE, 2013, pp. 611–615.

[11] B. Clerckx, Y. Mao, Z. Yang, M. Chen, A. Alkhateeb, L. Liu, M. Qiu, J. Yuan, V. W. S. Wong, and J. Montojo, "Multiple access techniques for intelligent and multifunctional 6g: Tutorial, survey, and outlook," *Proceedings of the IEEE*, vol. 112, no. 7, pp. 832–879, 2024.

[12] R. Prasad and T. Ojanpera, "An overview of CDMA evolution toward wideband CDMA," *IEEE communications Surveys*, vol. 1, no. 1, pp. 2–29, 1998.

[13] H.-H. Chen, H.-W. Chiu, and M. Guizani, "Orthogonal complementary codes for interference-free CDMA technologies," *IEEE Wireless Communications*, vol. 13, no. 1, pp. 68–79, 2006.

[14] H. Yin and S. Alamouti, "OFDMA: A broadband wireless access technology," in *2006 IEEE sarnoff symposium*. IEEE, 2006, pp. 1–4.

[15] H. Li, G. Ru, Y. Kim, and H. Liu, "OFDMA capacity analysis in MIMO channels," *IEEE transactions on information theory*, vol. 56, no. 9, pp. 4438–4446, 2010.

[16] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE communications surveys & tutorials*, vol. 20, no. 3, pp. 2294–2323, 2018.

[17] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE journal of selected topics in signal processing*, vol. 8, no. 5, pp. 742–758, 2014.

[18] A. J. Paulraj, D. A. Gore, R. U. Nabar, and H. Bolcskei, "An overview of MIMO communications-a key to gigabit wireless," *Proceedings of the IEEE*, vol. 92, no. 2, pp. 198–218, 2004.

[19] R. Knopp and P. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proceedings IEEE International Conference on Communications ICC '95*, vol. 1, 1995, pp. 331–335 vol.1.

[20] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1277–1294, 2002.

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

[22] H. Weingarten, Y. Steinberg, and S. S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE transactions on information theory*, vol. 52, no. 9, pp. 3936–3964, 2006.

[23] B. M. Hochwald, C. B. Peel, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-Part II: Perturbation," *IEEE Transactions on Communications*, vol. 53, no. 3, pp. 537–544, 2005.

[24] C. Windpassinger, R. F. Fischer, and J. B. Huber, "Lattice-reduction-aided broadcast precoding," *IEEE Transactions on Communications*, vol. 52, no. 12, pp. 2057–2060, 2004.

[25] A. Mishra, Y. Mao, O. Dizdar, and B. Clerckx, "Rate-splitting multiple access for 6G—Part I: Principles, applications and future works," *IEEE Communications Letters*, vol. 26, no. 10, pp. 2232–2236, 2022.

[26] Z. Pan, K.-K. Wong, and T.-S. Ng, "Generalized multiuser orthogonal space-division multiplexing," *IEEE Transactions on Wireless Communications*, vol. 3, no. 6, pp. 1969–1973, 2004.

[27] S. S. Christensen, R. Agarwal, E. De Carvalho, and J. M. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 4792–4799, 2008.

[28] B. Clerckx and C. Oestges, *MIMO wireless networks: channels, techniques and standards for multi-antenna, multi-user and multi-cell systems*. Academic Press, 2013.

[29] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE Journal on selected areas in Communications*, vol. 21, no. 5, pp. 684–702, 2003.

[30] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Transactions on information theory*, vol. 52, no. 11, pp. 5045–5060, 2006.

[31] B. Clerckx, H. Joudeh, C. Hao, M. Dai, and B. Rassouli, "Rate splitting for MIMO wireless networks: A promising PHY-layer strategy for LTE evolution," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 98–105, 2016.

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

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

[34] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 176–183, 2017.

[35] B. Clerckx, Y. Mao, R. Schober, E. A. Jorswieck, D. J. Love, J. Yuan, L. Hanzo, G. Y. Li, E. G. Larsson, and G. Caire, "Is NOMA efficient in multi-antenna networks? A critical look at next generation multiple access techniques," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 1310–1343, 2021.

[36] B. Clerckx, Y. Mao, R. Schober, and H. V. Poor, "Rate-splitting unifying SDMA, OMA, NOMA, and multicasting in MISO broadcast channel: A simple two-user rate analysis," *IEEE Wireless Communications Letters*, vol. 9, no. 3, pp. 349–353, 2019.

[37] B. Clerckx, Y. Mao, E. A. Jorswieck, J. Yuan, D. J. Love, E. Erkip, and D. Niyato, "A primer on rate-splitting multiple access: Tutorial, myths, and frequently asked questions," *IEEE Journal on Selected Areas in Communications*, 2023.

[38] Y. Mao, B. Clerckx, and V. O. Li, "Rate-splitting multiple access for downlink communication systems: bridging, generalizing, and outperforming SDMA and NOMA," *EURASIP journal on wireless communications and networking*, vol. 2018, no. 1, pp. 1–54, 2018.

[39] A. R. Flores, R. C. De Lamare, and B. Clerckx, "Tomlinson-Harashima Precoded Rate-Splitting With Stream Combiners for MU-MIMO Systems," *IEEE Transactions on Communications*, vol. 69, no. 6, pp. 3833–3845, 2021.

[40] C. Hao, Y. Wu, and B. Clerckx, "Rate Analysis of Two-Receiver MISO Broadcast Channel With Finite Rate Feedback: A Rate-Splitting Approach," *IEEE Transactions on Communications*, vol. 63, no. 9, pp. 3232–3246, 2015.

[41] A. Gholami Davoodi and S. A. Jafar, "Aligned Image Sets Under Channel Uncertainty: Settling Conjectures on the Collapse of Degrees of Freedom Under Finite Precision CSIT," *IEEE Transactions on Information Theory*, vol. 62, no. 10, pp. 5603–5618, Oct 2016.

[42] H. Joudeh and B. Clerckx, "Sum-rate maximization for linearly precoded downlink multiuser MISO systems with partial CSIT: A rate-splitting approach," *IEEE Transactions on Communications*, vol. 64, no. 11, pp. 4847–4861, 2016.

[43] Y. Mao, B. Clerckx, and V. O. Li, "Energy efficiency of rate-splitting multiple access, and performance benefits over SDMA and NOMA," in *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*. IEEE, 2018, pp. 1–5.

[44] B. Matthiesen, Y. Mao, A. Dekorsy, P. Popovski, and B. Clerckx, "Globally Optimal Spectrum- and Energy-Efficient Beamforming for Rate Splitting Multiple Access," *IEEE Transactions on Signal Processing*, vol. 70, pp. 5025–5040, 2022.

[45] Y. Mao, B. Clerckx, and V. O. Li, "Rate-splitting for multi-antenna non-orthogonal unicast and multicast transmission: Spectral and energy efficiency analysis," *IEEE Transactions on Communications*, vol. 67, no. 12, pp. 8754–8770, 2019.

[46] Y. Mao, B. Clerckx, J. Zhang, V. O. Li, and M. A. Arafah, "Max-min fairness of K-user cooperative rate-splitting in MISO broadcast channel with user relaying," *IEEE Transactions on Wireless Communications*, vol. 19, no. 10, pp. 6362–6376, 2020.

[47] H. Chen, D. Mi, B. Clerckx, Z. Chu, J. Shi, and P. Xiao, "Joint Power and Subcarrier Allocation Optimization for Multigroup Multicast Systems With Rate Splitting," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 2, pp. 2306–2310, Feb 2020.

[48] H. Joudeh and B. Clerckx, "Rate-splitting for max-min fair multigroup multicast beamforming in overloaded systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 11, pp. 7276–7289, 2017.

[49] O. Dizdar, Y. Mao, and B. Clerckx, "Rate-Splitting Multiple Access to Mitigate the Curse of Mobility in (Massive) MIMO Networks," *IEEE Transactions on Communications*, vol. 69, no. 10, pp. 6765–6780, 2021.

[50] G. Lu, L. Li, H. Tian, and F. Qian, "MMSE-Based Precoding for Rate Splitting Systems With Finite Feedback," *IEEE Communications Letters*, vol. 22, no. 3, pp. 642–645, March 2018.

[51] B. Rimoldi and R. Urbanke, "A rate-splitting approach to the Gaussian multiple-access channel," *IEEE Transactions on Information Theory*, vol. 42, no. 2, pp. 364–375, 1996.

[52] A. J. Grant, B. Rimoldi, R. L. Urbanke, and P. A. Whiting, "Rate-splitting multiple access for discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 873–890, 2001.

[53] Y. Zhu, X. Wang, Z. Zhang, X. Chen, and Y. Chen, "A rate-splitting non-orthogonal multiple access scheme for uplink transmission," in *2017 9th International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE, 2017, pp. 1–6.

[54] Z. Yang, M. Chen, W. Saad, W. Xu, and M. Shikh-Bahaei, "Sum-rate maximization of uplink rate splitting multiple access (RSMA) communication," *IEEE Transactions on Mobile Computing*, 2020.

[55] S. A. Tegos, P. D. Diamantoulakis, and G. K. Karagiannidis, "On the performance of uplink rate-splitting multiple access," *IEEE Communications Letters*, vol. 26, no. 3, pp. 523–527, 2022.

[56] H. Liu, T. A. Tsiftsis, K. J. Kim, K. S. Kwak, and H. V. Poor, "Rate splitting for uplink NOMA with enhanced fairness and outage performance," *IEEE Transactions on Wireless Communications*, vol. 19, no. 7, pp. 4657–4670, 2020.

[57] Y. Liu, B. Clerckx, and P. Popovski, "Network slicing for eMBB, URLLC, and mMTC: An uplink rate-splitting multiple access approach," *IEEE Transactions on Wireless Communications*, 2023.

[58] S. Schiessl, M. Skoglund, and J. Gross, "NOMA in the uplink: Delay analysis with imperfect CSI and finite-length coding," *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 3879–3893, 2020.

[59] J. Zeng, T. Lv, W. Ni, R. P. Liu, N. C. Beaulieu, and Y. J. Guo, "Ensuring max-min fairness of UL SIMO-NOMA: A rate splitting approach," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 11, pp. 11 080–11 093, 2019.

[60] H. Jiang, L. You, A. Elzanaty, J. Wang, W. Wang, X. Gao, and M.-S. Alouini, "Rate-splitting multiple access for uplink massive MIMO with electromagnetic exposure constraints," *IEEE journal on selected areas in communications*, vol. 41, no. 5, pp. 1383–1397, 2023.

[61] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[62] M. Médard, J. Huang, A. J. Goldsmith, S. P. Meyn, and T. P. Coleman, "Capacity of time-slotted ALOHA packetized multiple-access systems over the AWGN channel," *IEEE Transactions on Wireless Communications*, vol. 3, no. 2, pp. 486–499, 2004.

[63] J. Cao and E. M. Yeh, "Asymptotically optimal multiple-access communication via distributed rate splitting," *IEEE transactions on information theory*, vol. 53, no. 1, pp. 304–319, 2006.

[64] H. Joudeh and B. Clerckx, "Robust transmission in downlink multiuser MISO systems: A rate-splitting approach," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6227–6242, 2016.

[65] M. Medra and T. N. Davidson, "Robust downlink transmission: An offset-based single-rate-splitting approach," in *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2018, pp. 1–5.

[66] J. Zhang, B. Clerckx, J. Ge, and Y. Mao, "Cooperative rate splitting for MISO broadcast channel with user relaying, and performance benefits over

cooperative NOMA," *IEEE Signal Processing Letters*, vol. 26, no. 11, pp. 1678–1682, 2019.

[67] C. Hao, Y. Wu, and B. Clerckx, "Rate analysis of two-receiver MISO broadcast channel with finite rate feedback: A rate-splitting approach," *IEEE Transactions on Communications*, vol. 63, no. 9, pp. 3232–3246, 2015.

[68] E. Piovano and B. Clerckx, "Optimal DoF region of the K-user MISO BC with partial CSIT," *IEEE Communications Letters*, vol. 21, no. 11, pp. 2368–2371, 2017.

[69] M. Dai, B. Clerckx, D. Gesbert, and G. Caire, "A rate splitting strategy for massive MIMO with imperfect CSIT," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4611–4624, 2016.

[70] Y. Mao and B. Clerckx, "Beyond dirty paper coding for multi-antenna broadcast channel with partial CSIT: A rate-splitting approach," *IEEE Transactions on Communications*, vol. 68, no. 11, pp. 6775–6791, 2020.

[71] J. An, O. Dizdar, B. Clerckx, and W. Shin, "Rate-splitting multiple access for multi-antenna broadcast channel with imperfect CSIT and CSIR," *arXiv preprint arXiv:2102.08738*, 2021.

[72] X. Lyu, S. Aditya, J. Kim, and B. Clerckx, "A prototype implementation of rate splitting multiple access using software-defined radios," *arXiv preprint arXiv:2305.07361*, 2023.

[73] O. Dizdar, Y. Mao, W. Han, and B. Clerckx, "Rate-splitting multiple access for downlink multi-antenna communications: Physical layer design and link-level simulations," in *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*. IEEE, 2020, pp. 1–6.

[74] L. Yin, O. Dizdar, and B. Clerckx, "Rate-splitting multiple access for multi-group multicast cellular and satellite communications: PHY layer design and link-level simulations," in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2021, pp. 1–6.

[75] A. Mishra, Y. Mao, O. Dizdar, and B. Clerckx, "Rate-splitting multiple access for downlink multiuser MIMO: Precoder optimization and PHY-layer design," *IEEE Transactions on Communications*, vol. 70, no. 2, pp. 874–890, 2021.

[76] P. Mathys, "A class of codes for a t active users out of n multiple-access communication system," *IEEE Transactions on Information Theory*, vol. 36, no. 6, pp. 1206–1219, 1990.

[77] O. Abbasi and H. Yanikomeroglu, "Transmission scheme, detection and power allocation for uplink user cooperation with NOMA and RSMA," *IEEE Transactions on Wireless Communications*, vol. 22, no. 1, pp. 471–485, 2022.

[78] H. Kong, M. Lin, Z. Wang, J.-Y. Wang, W.-P. Zhu, and J. Wang, "Combined robust beamforming with uplink RSMA for multibeam satellite systems," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 9, pp. 10 167–10 172, 2022.

[79] Q. Sun, H. Liu, S. Yan, T. A. Tsiftsis, and J. Yuan, "Joint receive and passive beamforming optimization for RIS-assisted uplink RSMA systems," *IEEE Wireless Communications Letters*, 2023.

[80] "A mathematical theory of communication, author=Shannon, Claude Elwood," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[81] 5G Americas, "New services & applications with 5G ultra-reliable low latency communications," Nov. 2018. [Online]. Available: https://www.5gamericas.org/wp-content/uploads/2019/07/5G_Americas_URLLLC_White_Paper_Final___updateJW.pdf

[82] J. Lorca *et al.*, "Scenarios, KPIs, use cases and baseline system evaluation," E2E-aware Optimizations and advancements for Network Edge of 5G New Radio (ONE5G), Tech. Rep. ONE5G D2.1, Nov. 2017.

[83] 5G Americas, "Becoming 5G-advanced: the 3GPP 2025 roadmap," Dec. 2022. [Online]. Available: https://www.5gamericas.org/wp-content/uploads/2022/12/Becoming-5G-Advanced-the-3GPP-2025-Roadmap-InDesign.pdf

[84] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3309–3338, 2012.

[85] G. Ozcan and M. C. Gursoy, "Throughput of cognitive radio systems with finite blocklength codes," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 11, pp. 2541–2554, 2013.

[86] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4232–4265, 2014.

[87] W. Yang, G. Caire, G. Durisi, and Y. Polyanskiy, "Optimum power control at finite blocklength," *IEEE Transactions on Information Theory*, vol. 61, no. 9, pp. 4598–4615, 2015.

[88] Y. Hu, J. Gross, and A. Schmeink, "On the capacity of relaying with finite blocklength," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 3, pp. 1790–1794, 2015.

[89] L. Zhang and Y.-C. Liang, "Average throughput analysis and optimization in cooperative IoT networks with short packet communication," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 11 549–11 562, 2018.

[90] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *2012 Proceedings IEEE INFOCOM*. IEEE, 2012, pp. 2731–2735.

[91] R. Devassy, G. Durisi, G. C. Ferrante, O. Simeone, and E. Uysal-Biyikoglu, "Delay and peak-age violation probability in short-packet transmissions," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 2471–2475.

[92] R. Devassy, G. Durisi, G. C. Ferrante, O. Simeone, and E. Uysal, "Reliable transmission of short packets through queues and noisy channels under latency and peak-age violation guarantees," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 4, pp. 721–734, 2019.

[93] C. M. W. Basnayaka, D. N. K. Jayakody, T. D. P. Perera, and M. V. Ribeiro, "Age of information in an URLLC-enabled decode-and-forward wireless communication system," in *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*. IEEE, 2021, pp. 1–6.

[94] J. Scarlett, V. Y. Tan, and G. Durisi, "The dispersion of nearest-neighbor decoding for additive non-Gaussian channels," *IEEE Transactions on Information Theory*, vol. 63, no. 1, pp. 81–92, 2016.

[95] T. Cover, "Broadcast channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 2–14, 1972.

[96] M. Haghifam, M. R. Mili, B. Makki, M. Nasiri-Kenari, and T. Svensson, "Joint sum rate and error probability optimization: Finite blocklength analysis," *IEEE Wireless Communications Letters*, vol. 6, no. 6, pp. 726–729, 2017.

[97] Y. Yu, H. Chen, Y. Li, Z. Ding, and B. Vucetic, "On the performance of non-orthogonal multiple access in short-packet communications," *IEEE Communications Letters*, vol. 22, no. 3, pp. 590–593, 2017.

[98] X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen, and Z. Zhong, "Short-packet downlink transmission with non-orthogonal multiple access," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4550–4564, 2018.

[99] J. Chen, L. Zhang, Y.-C. Liang, and S. Ma, "Optimal resource allocation for multicarrier NOMA in short packet communications," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 2, pp. 2141–2156, 2019.

[100] Y. Xu, C. Shen, T.-H. Chang, S.-C. Lin, Y. Zhao, and G. Zhu, "Transmission energy minimization for heterogeneous low-latency NOMA downlink," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 1054–1069, 2019.

[101] S. Han, X. Xu, Z. Liu, P. Xiao, K. Moessner, X. Tao, and P. Zhang, "Energy-efficient short packet communications for uplink NOMA-based massive MTC networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 12, pp. 12 066–12 078, 2019.

[102] S. Schiessl, H. Al-Zubaidy, M. Skoglund, and J. Gross, "Delay performance of wireless communications with imperfect CSI and finite-length coding," *IEEE Transactions on Communications*, vol. 66, no. 12, pp. 6527–6541, 2018.

[103] S. Schiessl, J. Gross, M. Skoglund, and G. Caire, "Delay performance of the multiuser MISO downlink under imperfect CSI and finite-length coding," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 4, pp. 765–779, 2019.

[104] E. Dosit, M. Shehab, H. Alves, and M. Latva-aho, "Performance of Non-orthogonal Multiple Access under Finite Blocklength," *arXiv preprint arXiv:1902.09993*, 2019.

[105] N. P. Le and K. N. Le, "Performance analysis of NOMA short-packet communications with QoS-based SIC detecting order," *IEEE Wireless Communications Letters*, vol. 11, no. 3, pp. 617–621, 2021.

[106] Y. Xu, Y. Mao, O. Dizdar, and B. Clerckx, "Rate-splitting multiple access with finite blocklength for short-packet and low-latency downlink communications," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 11, pp. 12 333–12 337, 2022.

[107] ——, "Max-min fairness of rate-splitting multiple access with finite blocklength communications," *IEEE Transactions on Vehicular Technology*, 2022.

[108] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5g be?" *IEEE Journal on selected areas in communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[109] ETSI, "Mobile-edge computing – introductory technical white paper," https://portal.etsi.org/Portals/0/TBpages/MEC/Docs/Mobile-edge_Computing_-_Introductory_Technical_White_Paper_V1%2018-09-14.pdf, Sep. 2014, accessed at 15/09/2024.

[110] A. Al-Shuwaili and O. Simeone, "Energy-efficient resource allocation for mobile edge computing-based augmented reality applications," *IEEE Wireless Communications Letters*, vol. 6, no. 3, pp. 398–401, 2017.

[111] O. Salman, I. Elhajj, A. Kayssi, and A. Chehab, "Edge computing enabling the internet of things," in *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*. IEEE, 2015, pp. 603–608.

[112] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Joint allocation of computation and communication resources in multiuser mobile cloud computing," in *2013 IEEE 14th workshop on signal processing advances in wireless communications (SPAWC)*. IEEE, 2013, pp. 26–30.

[113] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM transactions on networking*, vol. 24, no. 5, pp. 2795–2808, 2015.

[114] X. Lyu, H. Tian, C. Sengul, and P. Zhang, "Multiuser joint task offloading and resource optimization in proximate clouds," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 4, pp. 3435–3447, 2016.

[115] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5g and beyond," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2347–2381, 2017.

[116] C. You, K. Huang, H. Chae, and B.-H. Kim, "Energy-Efficient Resource Allocation for Mobile-Edge Computation Offloading," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1397–1411, 2016.

[117] X. Li, R. Fan, H. Hu, N. Zhang, X. Chen, and A. Meng, "Energy-Efficient Resource Allocation for Mobile Edge Computing with Multiple Relays," *IEEE Internet of Things Journal*, vol. 9, no. 13, pp. 10 732–10 750, 2021.

[118] 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, 2019.

[119] S. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 721–742, 2016.

[120] W. U. Khan, F. Jameel, T. Ristaniemi, S. Khan, G. A. S. Sidhu, and J. Liu, "Joint Spectral and Energy Efficiency Optimization for Downlink NOMA Networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 2, pp. 645–656, 2019.

[121] W. U. Khan, J. Liu, F. Jameel, V. Sharma, R. Jäntti, and Z. Han, "Spectral Efficiency Optimization for Next Generation NOMA-Enabled IoT Networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 15 284–15 297, 2020.

[122] F. Wang, J. Xu, and Z. Ding, "Multi-Antenna NOMA for Computation Offloading in Multiuser Mobile Edge Computing Systems," *IEEE Transactions on Communications*, vol. 67, no. 3, pp. 2450–2463, 2019.

[123] Z. Ding, J. Xu, O. A. Dobre, and H. V. Poor, "Joint Power and Time Allocation for NOMA–MEC Offloading," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 6, pp. 6207–6211, 2019.

[124] J. Zhu, J. Wang, Y. Huang, F. Fang, K. Navaie, and Z. Ding, "Resource Allocation for Hybrid NOMA MEC Offloading," *IEEE Transactions on Wireless Communications*, vol. 19, no. 7, pp. 4964–4977, 2020.

[125] F. Fang, Y. Xu, Z. Ding, C. Shen, M. Peng, and G. K. Karagiannidis, "Optimal Resource Allocation for Delay Minimization in NOMA-MEC Networks," *IEEE Transactions on Communications*, vol. 68, no. 12, pp. 7867–7881, 2020.

[126] Z. Ding, D. W. K. Ng, R. Schober, and H. V. Poor, "Delay Minimization for NOMA-MEC Offloading," *IEEE Signal Processing Letters*, vol. 25, no. 12, pp. 1875–1879, 2018.

[127] Y. Ye, R. Q. Hu, G. Lu, and L. Shi, "Enhance Latency-Constrained Computation in MEC Networks Using Uplink NOMA," *IEEE Transactions on Communications*, vol. 68, no. 4, pp. 2409–2425, 2020.

[128] Z. Liu, Y. Zhu, Y. Hu, P. Sun, and A. Schmeink, "Reliability-Oriented Design Framework in NOMA-Assisted Mobile Edge Computing," *IEEE Access*, vol. 10, pp. 103 598–103 609, 2022.

[129] Y. Ye, G. Lu, R. Q. Hu, and L. Shi, "On the performance and optimization for mec networks using uplink noma," in *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2019, pp. 1–6.

[130] F. Jia, H. Zhang, H. Ji, and X. Li, "Distributed resource allocation and computation offloading scheme for cognitive mobile edge computing networks with noma," in *2018 IEEE/CIC International Conference on Communications in China (ICCC)*. IEEE, 2018, pp. 553–557.

[131] A. Chen, Z. Yang, B. Lyu, and B. Xu, "System delay minimization for noma-based cognitive mobile edge computing," *IEEE Access*, vol. 8, pp. 62 228–62 237, 2020.

[132] C. T. Nguyen, Q.-V. Pham, H.-G. T. Pham, N.-N. Dao, and W.-J. Hwang, "Computation offloading in cognitive radio noma-enabled multi-access edge computing systems," *IET Communications*, vol. 14, no. 19, pp. 3404–3409, 2020.

[133] M. Diamanti, C. Pelekis, E. E. Tsiropoulou, and S. Papavassiliou, "Delay Minimization for Rate-Splitting Multiple Access-Based Multi-Server MEC Offloading," *IEEE/ACM Transactions on Networking*, vol. 32, no. 2, pp. 1035–1047, 2024.

[134] F. Xiao, P. Chen, H. Wu, Y. Mao, and H. Liu, "Delay Minimization Using Hybrid RSMA-TDMA for Mobile Edge Computing," *Electronics*, vol. 12, no. 11, p. 2550, 2023.

[135] P. Chen, H. Liu, Y. Ye, L. Yang, K. J. Kim, and T. A. Tsiftsis, "Rate-Splitting Multiple Access Aided Mobile Edge Computing With Randomly Deployed Users," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 5, pp. 1549–1565, 2023.

[136] Y. Zhu, Y. Hu, A. Schmeink, and J. Gross, "Energy Minimization of Mobile Edge Computing Networks with HARQ in The Finite Blocklength Regime," *IEEE Transactions on Wireless Communications*, vol. 21, no. 9, pp. 7105–7120, 2022.

[137] X. Lai, T. Wu, C. Pan, L. Mai, and A. Nallanathan, "Short-Packet Edge Computing Networks With Execution Uncertainty," *IEEE Transactions on Green Communications and Networking*, 2024.

[138] J. Liu and Q. Zhang, "Offloading Schemes in Mobile Edge Computing for Ultra-Reliable Low Latency Communications," *Ieee Access*, vol. 6, pp. 12 825–12 837, 2018.

[139] M. Grant, S. Boyd, and Y. Ye, "CVX: Matlab software for disciplined convex programming," 2008.

[140] A. Ben-Tal and A. Nemirovski, *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM, 2001.

[141] J. Xu, O. Dizdar, and B. Clerckx, "Rate-splitting multiple access for short-packet uplink communications: A finite blocklength analysis," *IEEE Communications Letters*, vol. 27, no. 2, pp. 517–521, 2022.

[142] J. Xu and Y. Mao, "Max-Min Fairness for Uplink Rate-Splitting Multiple Access with Finite Blocklength," *arXiv preprint arXiv:2402.12066*, 2024.

[143] M. Grant, S. Boyd, and Y. Ye, "CVX: Matlab software for disciplined convex programming," 2009.

[144] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Transactions on information Theory*, vol. 55, no. 7, pp. 3051–3073, 2009.

[145] D. Seethaler, G. Matz, and F. Hlawatsch, "An efficient MMSE-based demodulator for MIMO bit-interleaved coded modulation," in *IEEE Global Telecommunications Conference, 2004. GLOBECOM'04.*, vol. 4. IEEE, 2004, pp. 2455–2459.

[146] I. Tal and A. Vardy, "List decoding of polar codes," *IEEE transactions on information theory*, vol. 61, no. 5, pp. 2213–2226, 2015.

[147] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE communications surveys & tutorials*, vol. 19, no. 3, pp. 1628–1656, 2017.

[148] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: design aspects, challenges, and future directions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 22–28, 2016.

[149] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint offloading and computing optimization in wireless powered mobile-edge computing systems," *IEEE transactions on wireless communications*, vol. 17, no. 3, pp. 1784–1797, 2017.

[150] S. Yang and Z. Lit, "A constant-gap result on the multi-antenna broadcast channels with linearly precoded rate splitting," in *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2018, pp. 1–5.

[151] Z. Li and S. Yang, "A linearly precoded rate splitting approach and its optimality for MIMO broadcast channels," in *2018 IEEE Information Theory Workshop (ITW)*. IEEE, 2018, pp. 1–5.

[152] L. Yin and B. Clerckx, "Rate-splitting multiple access for multigroup multicast and multibeam satellite systems," *IEEE Transactions on Communications*, vol. 69, no. 2, pp. 976–990, 2020.

[153] A. D. Zayas and P. Merino, "The 3GPP NB-IoT system architecture for the Internet of Things," in *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2017, pp. 277–282.

[154] H. Fu, S. Feng, W. Tang, and D. W. K. Ng, "Robust secure beamforming design for two-user downlink miso rate-splitting systems," *IEEE Transactions on Wireless Communications*, vol. 19, no. 12, pp. 8351–8365, 2020.

[155] P. Li, M. Chen, Y. Mao, Z. Yang, B. Clerckx, and M. Shikh-Bahaei, "Cooperative rate-splitting for secrecy sum-rate enhancement in multi-antenna broadcast channels," in *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*. IEEE, 2020, pp. 1–6.

[156] A. Mishra, Y. Mao, C. K. Thomas, L. Sanguinetti, and B. Clerckx, "Mitigating intra-cell pilot contamination in massive mimo: A rate splitting approach," *IEEE Transactions on Wireless Communications*, vol. 22, no. 5, pp. 3472–3487, 2022.