

Imperial College London
Department of Electrical and Electronic Engineering

Advancing Wireless Communications With Uplink Rate-Splitting Multiple Access

Yuanwen Liu
刘元文

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, 2025

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 conferences and 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

New services and demands for capacity and connectivity bring critical challenges to current networks and how to allocate resources more efficiently has become an important problem. This calls for reconsidering multiple access (MA) techniques in future networks. As a general and efficient MA technique, Rate-Splitting Multiple Access (RSMA) has attracted attention from academia and industry. Most of the works of RSMA are related to downlink, while exploiting uplink RSMA still needs further exploration. This thesis focused on integrating uplink RSMA with emerging techniques and enhancing the reliability of uplink RSMA.

The applications of uplink RSMA in two new techniques are investigated, namely network slicing and semantic communications. First, RSMA with network slicing is studied. Enhanced Mobile Broadband (eMBB), ultra-reliable and low-latency communications (URLLC) and massive machine type communications (mMTC) are three generic services in future networks, and network slicing is a technique to satisfy the requirements of different services. An uplink RSMA approach for network slicing is investigated to allocate spectrum resources efficiently while guaranteeing different requirements of users.

Then, the application of RSMA in the coexistence of semantic communications and bit communications is studied. Semantic communication is efficient for users who only care about the semantic meaning, but it will not completely replace bit communication, so there will be coexistence between semantic communications and bit communications in future. An uplink RSMA scheme is investigated to allocate resources between these two types of communications efficiently.

Next, considering the high reliability requirements in future networks, integrating RSMA and hybrid automatic repeat request (HARQ) is investigated. Conventional HARQ that retransmitting all the failed messages is not suitable for RSMA, since it may increase the complexity. A HARQ scheme tailored for RSMA is studied, and the results show that RSMA with HARQ has promising reliability and energy efficiency.

Acknowledgements

First, I would like to express my sincere gratitude and appreciation to my supervisor, Prof. Bruno Clerckx, for his instructions and guidance during my PhD study.

Then, I would like to appreciate all my colleagues and friends. Thank you for the discussions about research, and this is really a big support for my PhD journey. I would also like to thank you for sharing the happy time with me, and I will never forget our stories in those tea breaks, werewolf, card games and also the time we spent in Austria. I will forever cherish those moments that we learned, explored and had tons of fun.

Thanks to Professor Petar Popovski. Thank you for your guidance during our corporations, and your insightful comments support me a lot.

Thanks to my best friend Yutong. Thank you for accompanying me just as we were teenagers. I can always know the different perspectives from you and thank you for all your supports during my PhD. I always wish you the best.

Thanks to my beloved. Thank you for your continued love and support. You always give me interesting points of view. I am very happy to talk with you about history, philosophy, literature and everything. You have enriched my life outside the research.

Thanks to my Jellicats. Whenever I come home and see you, I feel warm and motivated.

Lastly, but most importantly, I thank my family. Thank you for your unconditional love, sacrifices, understanding and encouragement. In these years I am abroad, when I think of you I feel like I can do anything as long as you are behind me. My courage and vigor are inseparable from your consistent support, and they are extremely important for me to complete my PhD.

Yuanwen Liu

刘元文

London, January 2025.

Abbreviations

5G Fifth Generation

6G Sixth Generation

ACK Acknowledgment

AI Artificial Intelligence

AoI Age of Information

AoII Age of Incorrect Information

ARQ Automatic Repeat Request

B5G Beyond Fifth-Generation

BS Base Station

CC Chase Combining

CDMA Code Division Multiple Access

CSIT Channel State Information at Transmitter

CSI Channel State Information%acrodpc[DPC]Dirty Paper Coding

DL Deep Learning

eMBB enhanced Mobile Broadband

FDMA Frequency Division Multiple Access

FEC Forward Error Correction

HARQ Hybrid Automatic Repeat Request

i.i.d independent and identically distributed

IR incremental redundancy

MA Multiple Access

MAC Multiple Access Channel

MIMO Multiple-Input Multiple-Output

mMTC massive Machine Type Communications

MRC maximal ratio combining

NACK Negative Acknowledgment

NLP Natural Language Processing

NOMA Non-Orthogonal Multiple Access

OFDMA Orthogonal Frequency Division Multiple Access

OMA Orthogonal Multiple Access

PDF Probability Density Function

QoS Quality-of-Service

RSMA Rate Splitting Multiple Access

SCA Successive Convex Approximation

SC Superposition Coding

SDMA Space Division Multiple Access

SE Spectral Efficiency

SIC Successive Interference Cancellation

SINR Signal to Interference plus Noise Ratio

SISO Single-Input Single-Output

SNR Signal-to-Noise Ratio

SQP Sequential Quadratic Programming

TDMA Time Division Multiple Access

URLLC Ultra-Reliable and Low-Latency Communications

Notations

Constants

e Euler's number $\simeq 2.71828\dots$

j Imaginary unit $= \sqrt{-1}$

π Archimedes' constant $\simeq 3.14159\dots$

Objects

a Scalar

$\mathbf{a}$ Column vector

$\mathbf{A}$ Matrix

$\mathcal{A}$ Finite set

Sets

$\mathbb{N}$ Natural numbers

$\mathbb{R}$ Real numbers

$\mathbb{C}$ Complex numbers

Operations

$(\cdot)^\top$ Transpose

$(\cdot)^H$ Hermitian (conjugate transpose)

$|\cdot|$ Absolute value of a complex number

$\ln(\cdot)$ Natural logarithm of a real number

$\exp(\cdot)$ Exponential of a scalar or square matrix

$\mathbb{E}[\cdot]$ Expectation operator

$\Gamma(s, x)$ The upper incomplete gamma function, $\Gamma(s, x) = \int_x^\infty t^{s-1} e^{-t} dt$

$Q(x)$ Gaussian Q function, $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$

$\Pr(A)$ The probability of event A

$\max\{a\}$ The maximum value of a

$\cup$ The union of the sets

Distributions

$\sim$ Follows a distribution

$\mathcal{CN}(\delta, \sigma^2)$ Circularly symmetric complex Gaussian with mean δ and standard variance σ

Possion(λ) A Possion distribution with mean λ

Superscripts

$(\cdot)^{(i)}$ i -th iterated value

$(\cdot)^\star$ Stationary point

Contents

Declaration of Originality	1
Copyright Declaration	3
Abstract	5
Acknowledgements	7
Abbreviations	9
Notations	13
1 Introduction	25
1.1 Motivation	26
1.2 Rate-Splitting Multiple Access: A Promising Strategy for Future Com- munications	27
1.3 Organization and Contributions	30
1.4 Publications	32
2 Preliminaries and Literature Review	34
2.1 Fundamentals of Uplink RSMA	34
2.2 Network Slicing and Models of eMBB, URLLC and mMTC	40
2.2.1 eMBB	45
2.2.2 URLLC	47
2.2.3 mMTC	48

2.3	Semantic communications and Its Metric	50
2.3.1	Semantic Communication Metric	54
2.3.2	Data Regression Method For Semantic Rate Approximation	56
2.4	Hybrid-ARQ	57
3	RSMA Facilitating Network Slicing	60
3.1	Introduction	61
3.2	System Model	63
3.3	Slicing for eMBB and URLLC	65
3.3.1	OMA for eMBB and URLLC	66
3.3.2	NOMA for eMBB and URLLC	66
3.3.3	RSMA for eMBB and URLLC	68
3.4	Slicing for eMBB and mMTC	72
3.4.1	OMA for eMBB and mMTC	72
3.4.2	NOMA for eMBB and mMTC	72
3.4.3	RSMA for eMBB and mMTC	74
3.5	Numerical Results	77
3.5.1	eMBB and URLLC Coexistence	77
3.5.2	eMBB and mMTC Coexistence	82
3.6	Summary	86
4	RSMA for Coexistence of Semantic and Bit Communications	88
4.1	Introduction	89
4.2	System Model	91
4.2.1	FDMA	93
4.2.2	NOMA	98
4.2.3	RSMA	102
4.3	Numerical Results	108
4.3.1	The rate regions for different numbers of semantic users	108

4.3.2	Comparing RSMA behaviour for bit communications and for coexistence	111
4.3.3	The advantages of RSMA with multiple semantic users	112
4.3.4	The advantages of RSMA in high accuracy applications	115
4.4	Summary	116
5	Uplink RSMA with Hybrid ARQ	118
5.1	Introduction	118
5.2	System Model	123
5.3	HARQ design for RSMA	125
5.3.1	How α Determines Which Stream Needs Retransmissions . . .	125
5.3.2	Chase Combining	133
5.3.3	Incremental Redundancy	140
5.4	HARQ Scheme for FDMA and NOMA	141
5.5	Numerical Results	143
5.6	Summary	150
6	Conclusion	152
6.1	Summary of Thesis Achievements	152
6.2	Future Work	154
	References	155
1	Appendix A	170
2	Appendix B	171

List of Tables

2.1	Requirements of eMBB, URLLC and mMTC applications	41
5.1	Choosing α according to α_h and α_l	130

List of Figures

2.1	An illustration of the two-user rate region of MAC.	38
2.2	Transceiver architecture of K -user uplink RSMA.	39
2.3	Illustration of network slicing. (a) is the OMA scenario, and each service is allocated isolated resources, but the users of the same services can share the same time and frequency resources, i.e. one Ultra-Reliable and Low-Latency Communications (URLLC) user can occupy an isolated resource denoted by light blue, and multiple URLLC can also share the same resource which is denoted by dark blue. (b) is the general non-orthogonal slicing scenario, i.e., NOMA and RSMA. mMTC and eMBB, eMBB and URLLC can share the resources, and the sharing parts are denoted with black boxes.	43
2.4	The illustration of the framework of DeepSC.	54
2.5	The logistic regression for approximating $\varepsilon(\rho, K)$, ...	56
3.1	The considered scenario that devices from three services transmit to a common BS in the uplink.	64
3.2	An illustration of rate regions of TDMA, FDMA, NOMA and RSMA for a generic two-user case. The average channel gain of two users is 20 dB.	71
3.3	Rate region (r_B^{sum}, r_U^{sum}) , where r_B^{sum} , r_U^{sum} are the sum-rate of eMBB and URLLC, respectively. $\Gamma_B = 10$ dB, $\Gamma_U = 20$ dB, $F = 10$, $\epsilon_B = 10^{-3}$, $\epsilon_U = 10^{-5}$	79

3.4	Rate region (r_B^{sum}, r_U^{sum}) , where r_B^{sum}, r_U^{sum} are the sum-rate of eMBB and URLLC, respectively. $\Gamma_B=20$ dB, $\Gamma_U=10$ dB, $F=10$, $\epsilon_B=10^{-3}$, $\epsilon_U=10^{-5}$	80
3.5	r_U^{sum} versus power factor β . $\Gamma_B=10$ dB, $\Gamma_U=20$ dB, $F=10$, $\epsilon_B=10^{-3}$, $\epsilon_U=10^{-5}$	81
3.6	r_U^{sum} versus power factor β . $\Gamma_B=20$ dB, $\Gamma_U=10$ dB, $\epsilon_B=10^{-3}$, $\epsilon_U=10^{-5}$	81
3.7	Two illustrations of rate region	83
3.8	Achievable pairs (r_B, λ_M) , where r_B, λ_M are the rate of eMBB and the arrival rate of mMTC traffic, respectively. $\Gamma_B=20$ dB, $\Gamma_M=5$ dB, $r_M=0.04$ (bits/s/Hz), $\epsilon_B=10^{-3}$, $\epsilon_M=10^{-1}$	84
3.9	Arrival rate λ_M versus power factor β . $\Gamma_B=20$ dB, $\Gamma_M=5$ dB, $r_M=0.04$ (bits/s/Hz), $\epsilon_B=10^{-3}$, $\epsilon_M=10^{-1}$	86
4.1	An illustration of semantic users and bit user coexistence.	92
4.2	The illustration of FDMA.	93
4.3	The illustration of NOMA.	98
4.4	The decoding order of NOMA.	99
4.5	The illustration of RSMA.	103
4.6	The decoding order of RSMA.	104
4.7	The rate region of the coexistence of one semantic user and one bit user. X-axis denotes the semantic rate of semantic users and y-axis denotes the bit rate of the bit user. Case 1: $d_s=20$ meters, $d_b=30$ meters; case 2: $d_s=30$ meters, $d_b=30$ meters; case 3: $d_s=40$ meters, $d_b=30$ meters.	110
4.8	Power allocation strategies	112
4.9	The rate region of the coexistence of four semantic users and one bit user. X-axis denotes the semantic rate of semantic users and y-axis denotes the bit rate of the bit user. Case 1: $d_s=20$ meters, $d_b=30$ meters; case 2: $d_s=30$ meters, $d_b=30$ meters; case 3: $d_s=40$ meters, $d_b=30$ meters.	114

4.10	The number of semantic users versus the achievable bit rate.	115
4.11	The rate regions of $S^{th}=0.7,0.8,0.9$, $N_s=4$	116
5.1	The retransmission schemes of FDMA, NOMA and RSMA	120
5.2	An illustration of capacity region for two users.	128
5.3	A 3-user case	130
5.4	The rate region of 3-user case	132
5.5	The number of users splitting messages.	133
5.6	The situation that only s_2 needs retransmission. BS fails to decode s_2 and $s_{1,2}$, but only asks user 2 to retransmit s_2 . Then, BS uses two interfered copies of s_2 to decode it, and then continues decoding $s_{1,2}$	135
5.7	The situation that both streams need retransmission. BS fails to decode $s_{1,1}$ and asks both user 1 and user 2 to retransmit. Then, BS uses two interfered copies of $s_{1,1}$ and s_2 to decode them and then continues decoding $s_{1,2}$	136
5.8	Error probabilities of two users in CC. The average channel gains for user 1 and user 2 are 20 dB and 15 dB, respectively. Retransmission times $L=2$	145
5.9	Average power consumption of each packet in CC with 2 retransmissions.	145
5.10	Error probabilities of two users in IR. The average channel gains for user 1 and user 2 are 20 dB and 15 dB, respectively. Retransmission times $L=2$	146
5.11	Average power consumption of each packet in IR with 2 retransmissions.	147
5.12	Error probabilities of two users in CC. The average channel gains for user 1 and user 2 are 20 dB and 15 dB, respectively. Retransmission times $L=4$	148
5.13	Average power consumption of each packet in CC with 4 retransmissions.	149

5.14	Error probabilities of two users in IR. The average channel gains for user 1 and user 2 are 20 dB and 15 dB, respectively. Retransmission times $L=4$	150
5.15	Average power consumption of each packet in IR with 4 retransmissions.	151

Chapter 1

Introduction

Lifestyle and society have changed dramatically in the last decades, and they drive the need for network development [1]. We witness that Fifth Generation (5G) systems are becoming a commercial reality, and now Beyond Fifth-Generation (B5G) and Sixth Generation (6G) spark the interest of researchers. Future networks aim to provide services with enhanced Mobile Broadband (eMBB), URLLC, and massive Machine Type Communications (mMTC). To fulfill the requirements of services, the demands for capacity and connectivity increase unprecedentedly, which poses critical challenges to the current networks, and this calls for a more efficient resource allocation strategy. On the other hand, as Artificial Intelligence (AI) develops and becomes prevalent, it also brings new techniques to communications, such as semantic communications. Semantic communication is a promising solution to use the resources more efficiently for users who only care about semantic information. Thus, it raises the question of how to accommodate bit communications and semantic communications simultaneously. Therefore, it is important to reconsider the Multiple Access (MA) techniques to allocate resources more efficiently to alleviate the conflict between high demands and limited resources for users with different requirements and different types of communications.

1.1 Motivation

eMBB, URLLC and mMTC are three important services in future networks. These three services have different requirements and how to accommodate them in the communication networks is worthy of studying. Network slicing is a promising technique to satisfy different service requirements [2, 3], and it is usually done orthogonally and the users are allocated isolated resources. However, as the number of users grows, it is impossible for every user to use dedicated resources. Therefore, a more efficient MA scheme needs to be investigated.

The emerging services also bring the challenges of high demands of capacity and connectivity. As AI develops, semantic communication has attracted attention from both academia and industry [4–7]. Semantic communication aims to only transmit the semantic meaning of a message and this is favourable for the users who only care about the semantic meaning. Semantic communication can improve efficiency, but some users will still require the original messages, i.e. collecting raw data for analysis. This gives the motivation to investigate a suitable MA which can accommodate semantic users and bit users efficiently.

In this thesis, we focus on the non-orthogonal MA scheme, which uses resources more efficiently than the orthogonal one, specifically, Rate Splitting Multiple Access (RSMA). RSMA has attracted widespread attention and shown its advantages in spectrum efficiency in various scenarios [8–11]. Besides spectrum efficiency, reliability is another important indicator of future networks. Hybrid Automatic Repeat Request (HARQ) is a well-established mechanism to improve the reliability [12], and it is based on Forward Error Correction (FEC) and Automatic Repeat Request (ARQ) [13]. However, considering the unique characteristics of RSMA that users split its message into multiple parts, applying HARQ directly to RSMA is not the best solution to improve the reliability and it may also bring additional complexity. Therefore,

this also motivates us to investigate the HARQ scheme tailored for uplink RSMA.

1.2 Rate-Splitting Multiple Access: A Promising Strategy for Future Communications

To accommodate the various services in B5G and 6G with high capacity and low latency demands, MA techniques need to be reconsidered. Usually, MA techniques can be divided into two groups according to their interference management methods [14]: Orthogonal Multiple Access (OMA) and Non-Orthogonal Multiple Access (NOMA). OMA allocates isolated resources to users, while NOMA allows users to share the same resources.

In past generations, OMA has been used in cellular networks. Users occupy orthogonal radio resources in different domains: time domain, frequency domain, code domain and the combination of time and frequency domain, namely Time Division Multiple Access (TDMA), Frequency Division Multiple Access (FDMA), Code Division Multiple Access (CDMA) and Orthogonal Frequency Division Multiple Access (OFDMA), respectively [15]. Although users do not interfere with each other in OMA, it is not efficient as the number of users grows. This drove the exploration of MA techniques which enhance Spectral Efficiency (SE) [12]. As Multiple-Input Multiple-Output (MIMO) became an important role in wireless communications because it supports the exploding number of users, naturally the spatial dimension is exploited and it brought Space Division Multiple Access (SDMA) [16].

Different from OMA that users are interference-free, SDMA allows the occurrence of interference. In SDMA, users share the same time-frequency resources but they are separated through spatial beams. By designing the precoder (or beamformer) wisely, the interference between users can be mitigated significantly [17]. However,

SDMA still has its limitations. First, the performance SDMA degraded quickly in overloaded regime, i.e., the number of users is more than the number of antennas of the transmitter [18]. Second, SDMA needs careful scheduling to pair users with nearly orthogonal channels and similar strength. The complexity of the algorithm of scheduling increases rapidly even if there exists some suboptimal and low-complexity scheduling and user-pairing algorithm [19]. Third, SDMA is sensitive to the accuracy of Channel State Information (CSI) and imperfect Channel State Information at Transmitter (CSIT) can lead to the decrease in performance of SDMA [20].

Another non-orthogonal MA technique that has attracted attention is NOMA, and it separates users either in the power domain, namely power-domain NOMA, or code-domain, namely code-domain NOMA. In this work, we focus on power-domain NOMA¹. The transmitter uses Superposition Coding (SC) to superpose the message of users in the power domain and at the receiver side the messages are decoded by Successive Interference Cancellation (SIC). NOMA improves SE because it can leverage the channel strength differences, but it also has limitations. First, the performance of NOMA decreases when the channel strengths of users are similar. Because of using SIC to decode the messages, the user decoded first will tolerate much interference from other users, which may cause the decoding failure. Second, from a user fairness perspective, NOMA is not an optimal choice [21, 22]. Third, in the Multiple Access Channel (MAC), NOMA with time sharing can achieve the whole capacity region of MAC, but it needs multiple time slots [12] and induces communication overhead and latency, which brings additional complexity [23].

SDMA and NOMA can be seen as two extreme MA techniques. SDMA treats the interference between users as noise, while NOMA fully decodes the interference. Their limitations come from this inflexible interference management, and it calls for the MA technique which has a more flexible interference management. Therefore, as a

¹In this thesis, power-domain NOMA will be referred to NOMA.

more flexible MA technique, RSMA gains considerable attention from both academia and industry.

In the downlink, the universality of RSMA is shown in [17]. RSMA can softly bridge SDMA and NOMA and outperform them. [24] show the flexibility of RSMA, which can adapt to varying network loads and channel directions and strengths. [25, 26] show that RSMA is robust to imperfect CSI. [8–11] show that RSMA can enhance spectrum efficiency in various scenarios such as MIMO, unicast, multicast, satellite-terrestrial networks, beyond diagonal reconfigurable intelligent surface. [9, 27] show the advantage of RSMA from an energy efficiency perspective. [28–30] show that RSMA can enhance the Quality-of-Service (QoS) and user fairness. RSMA also shows its promising performances in the finite blocklength regime [31]. Since RSMA can achieve the same rate as SDMA and NOMA with a shorter blocklength, it has the potential to be used in low-latency applications. [32, 33] investigated the trade-off between spectral efficiency and secrecy performance, and shows that RSMA still has rate benefits while guaranteeing secrecy. [34] proposes the low-complexity design for RSMA with considering the practical receivers implementation and shows that RSMA still holds advantages over SDMA. The prototype of downlink RSMA is built and its performance is evaluated [35] and RSMA shows better performance than SDMA and NOMA.

Uplink RSMA is first proposed in [36]. In the uplink, RSMA also shows its advantages in improving spectrum efficiency [23, 37], enhancing user fairness [21, 22, 38, 39] in MIMO, finite blocklength regime. It can also simplify the implementation by avoiding complex user pairing [40], and increase the connectivity in semi-grant-free transmission [41]. [42] shows that uplink RSMA has the potential to be applied in low latency applications with low complexity. Therefore, RSMA is a promising technique for any service which needs grant-free and is required intermittently, i.e., URLLC and mMTC [43].

Since RSMA has shown numerous good properties, it is worth investigating RSMA

further. Most of the works about RSMA are in downlink, but uplink RSMA is quite different from downlink RSMA, so it cannot be simply extended from those downlink works, and this work will focus on the uplink RSMA. The main objective of this thesis is to investigate allocating resources efficiently between various services (eMBB, URLLC and mMTC) and users requiring different types of communications (bit communications and semantic communications) with RSMA, which are the key enablers of future networks. This thesis also investigates a HARQ scheme tailored for RSMA to improve reliability, which is an important indicator for future networks.

1.3 Organization and Contributions

Future networks are expected to provide more services and various types of communications, i.e. semantic communications, and allocate resources more efficiently between different services. Although previous works have shown the potential of uplink RSMA, most of them focused on a specific service and did not consider the resource allocation between services or between different type of communications. This motivates us to investigate allocating resources between various services (eMBB, URLLC and mMTC) and users requiring different types of communications (bit communications and semantic communications) with RSMA. Reliability is also an important indicator in future networks, so we investigate a HARQ scheme tailored for RSMA, which was not studied before.

A summary of each chapter, highlighting the main contributions, is given as follows.

- **Chapter 2 Preliminaries:** We review the fundamentals and related literature of this thesis by introducing the the principles of uplink RSMA, system model of network slicing, semantic communication and schemes of HARQ for uplink RSMA.

- **Chapter 3 An Uplink RSMA Approach for Network Slicing:** For the *first time*, an overall uplink RSMA scheme is investigated for network slicing scenario. Two coexistence scenarios are considered: eMBB and URLLC coexistence and eMBB and mMTC coexistence. For the eMBB and URLLC coexistence scenario, the achievable rate pairs of eMBB and URLLC of different MA schemes are characterized. For the eMBB and mMTC coexistence scenario, the achievable rate of eMBB and arrival rate of mMTC pairs of different MA schemes are characterized. The simulated results show that RSMA has promising performances. A detailed analysis of where the gains are from and the suitable scenarios of RSMA is provided. The numerical results and analysis show that RSMA improves the spectrum efficiency in the network slicing scenario in which users require various services with different requirements, and it has the potential to be applied in future networks.

- **Chapter 4 RSMA for Coexistence of Semantic and Bit Communications:** We investigate the MA schemes for semantic and bit communications coexistence scenario, and *for the first time* uplink RSMA is studied in this coexistence scenario. The RSMA scheme is designed to maximize the achievable rate of bit users while ensuring the understandability and transmission efficiency of semantic users. To solve this optimization problem, a Successive Convex Approximation (SCA) algorithm is proposed. The numerical results show that RSMA allocates resources efficiently in the coexistence scenario. It always outperforms NOMA and outperforms OMA in most situations. The differences of RSMA behaviours between coexistence scenarios and bit communications scenarios are summarized and analyzed. These results show the potential of RSMA in future networks.

- **Chapter 5 RSMA with HARQ:** A HARQ scheme tailored for uplink RSMA is investigated *for the first time*. HARQ is an important mechanism to improve

the reliability. Since in RSMA users split their messages, if applying conventional HARQ directly to retransmit all the failed streams, the complexity will increase. We show that the HARQ scheme we design for RSMA will not increase the complexity and the analysis is provided. This HARQ scheme is designed to minimize the error probabilities of the users and the related expressions of error probabilities are given. Numerical results show that RSMA with HARQ has promising reliability and consumes less power compared to NOMA and OMA.

- **Chapter 6 Conclusion:** We conclude the thesis by summarizing the achievements and discussing possible directions for future works.

1.4 Publications

The work in this thesis has resulted in several papers, that have been published, accepted or under review, or are still under preparation.

Under Review

- **Y. Liu** and B. Clerckx, “Rate-Splitting Multiple Access for Coexistence of Semantic and Bit Communications,” *submitted to IEEE Transactions on Wireless Communications*.

Published Journals

- **Y. Liu**, B. Clerckx and P. Popovski, “Network Slicing for eMBB, URLLC, and mMTC: An Uplink Rate-Splitting Multiple Access Approach,” in *IEEE Transactions on Wireless Communications*, vol. 23, no. 3, pp. 2140-2152, March 2024, doi: 10.1109/TWC.2023.3295804.

- **Y. Liu**, B. Clerckx and P. Popovski, "Performance Analysis of Uplink Rate-Splitting Multiple Access With Hybrid ARQ," in IEEE Transactions on Wireless Communications, vol. 23, no. 10, pp. 14201-14214, Oct. 2024, doi: 10.1109/TWC.2024.3410315.

Chapter 2

Preliminaries and Literature Review

In this chapter, the background knowledge and state-of-the-art works covered in this thesis are presented, including fundamentals of uplink RSMA, network slicing, semantic communications, and HARQ.

2.1 Fundamentals of Uplink RSMA

According to the discussions in Section 1.2, OMA allocates dedicated resources to users, i.e. TDMA allocates different time slots to different users, FDMA allocates different frequency resources to different users. While NOMA allows users to share resources, and the transmitter sends the superposition of messages and the receiver uses SIC to decide messages. Although they are two different types of MA techniques, they all transmit messages directly. Instead of transmitting the whole message, RSMA splits its messages into two parts and sends the superposition of the split parts. In downlink, Base Station (BS) splits the messages of each user into two parts, one common part and one private part. Then, all the common parts are encoded

into a common stream, and each private part is encoded into private streams. BS sends the superposition of common stream and private streams, and users decode the common stream first and then decode their private streams [17]. Unlike downlink RSMA, uplink RSMA does not have the definition of common part and private part. Users split their messages into two parts and send the superposition of them, and then the BS uses SIC to decode the superposed streams, and it has been shown that all the points of the rate region can be achieved by uplink RSMA [36]. Since BS needs to decode all the received streams, the decoding order in uplink cannot be directly extended from the downlink one, so the main challenges of uplink RSMA come from the decoding order design and the resource allocation.

In this section, we mainly introduce the fundamentals of uplink RSMA, and we focus on the multi-user Single-Input Single-Output (SISO) system, where K single-antenna users transmit to a single-antenna BS. We first introduce a two-user case to provide insight into the benefits of uplink RSMA. According to [36], the benefit of uplink RSMA is it can achieve all the points of the rate region by splitting the message of one of the users. First, we need to define the concept of rate region. The rate region is the set of all pairs (R_1, R_2) such that simultaneously user 1 and user 2 can reliably communicate at rate R_1 and R_2 , respectively [44]. For a two-user case in uplink, the rate region is shown in Fig. 2.1, which is a pentagon, and segment AB contains all the 'optimal' operating points. Point A and point B can be achieved by NOMA without time-sharing. For example, if BS receives s_1 from user 1 and s_2 from user 2, NOMA can achieve point A when the decoding order is s_1, s_2 ; and it can achieve point B when the decoding order is s_2, s_1 . However, neither of decoding order can enable NOMA without time-sharing to achieve other points on the segment AB . If the objective is to maximize the sum rate, either achieving point A or point B can reach the goal. However, some systems may want to ensure user fairness, i.e. operating at the points of AB other than point A and point B , and RSMA can achieve those points. We note that although NOMA with time-sharing can achieve all the points of AB , it induces communication

overhead and requires synchronization among users, so it can be difficult to apply in practice [12] and we only consider NOMA without time-sharing in this thesis.

After clarifying the definition of rate region and why points on segment AB except for point A and point B cannot be achieved by NOMA, then, we introduce how RSMA can achieve those points. [36] shows that by splitting the message of one of the users into two parts, all the points of rate region can be achieved. The detailed proof is provided in [36] for interested readers, and here we explain the sufficiency of this setup in a more intuitive way. First, we consider that only one user splits its message. Without the loss of generality, we assume that user 1 splits its message into two parts and then encodes them to $s_{1,1}$ and $s_{1,2}$, and user 2 transmits its encoded message s_2 directly. This can be seen as adding a virtual user with the same channel as user 1, and it is similar to a three-user case in NOMA which two users have the same channels. For any type of message to be transmitted, i.e., text, image, and video, user 1 first splits it into two parts. Since uplink RSMA does not have a common part and a private part, the split parts are equivalent and they do not need to be distinguished specifically from a content perspective. Then, user 1 encodes these two parts into bit streams, and interleaves and modulates bit streams to symbol streams $s_{1,1}$ and $s_{1,2}$ [45, 46]. User 2 uses the same procedure and transmits s_2 . The received signal at BS is

$$y = \sqrt{P_{1,1}}h_1s_{1,1} + \sqrt{P_{1,2}}h_1s_{1,2} + \sqrt{P_2}h_2s_2 + n, \quad (2.1)$$

where $P_{1,1}$ and $P_{1,2}$ are the power allocated to $s_{1,1}$ and $s_{1,2}$, respectively, h_1 and h_2 denote the channel of user 1-BS and the channel of user 2-BS, respectively, and n is the additive Gaussian noise with power of σ^2 . Instead of decoding $s_{1,1}$ and $s_{1,2}$ successively, the BS decodes the received streams in the order of $s_{1,1}$, s_2 and $s_{1,2}$. The rationale is that decoding $s_{1,1}$ and $s_{1,2}$ successively is equivalent to not splitting according to the chain rule. For some situations, splitting could be unnecessary, and these situations can be seen as either $P_{1,1} = 0$ or $P_{1,2} = 0$. Therefore, $s_{1,1}$, s_2 and $s_{1,2}$ is

the most general decoding order for a two-user case. The Signal to Interference plus Noise Ratio (SINR) of $s_{1,1}$, s_2 and $s_{1,2}$ can be represented as

$$\rho_{1,1} = \frac{P_{1,1}|h_1|^2}{P_{1,2}|h_1|^2 + P_2|h_2|^2 + \sigma^2}, \quad (2.2)$$

$$\rho_2 = \frac{P_2|h_2|^2}{P_{1,2}|h_1|^2 + \sigma^2} \quad (2.3)$$

and

$$\rho_{1,2} = \frac{P_{1,2}|h_1|^2}{\sigma^2}, \quad (2.4)$$

respectively. The achievable rates of $s_{1,1}$, s_2 and $s_{1,2}$ are

$$r_{1,1} = \log_2(1 + \rho_{1,1}), \quad (2.5)$$

$$r_2 = \log_2(1 + \rho_2) \quad (2.6)$$

and

$$r_{1,2} = \log_2(1 + \rho_{1,2}). \quad (2.7)$$

The achievable rate of user 1 is $r_1 = r_{1,1} + r_{1,2}$. This decoding order enables RSMA to achieve all the points of the rate region, which is shown Fig. 2.1.

The above discussion shows that splitting one user enable all the points of rate region to be achieved. Then, we explain that it is not necessary to let all users split their messages. We assume that two users split messages can achieve a 'better' point with the same transmission power which splitting one user cannot, and we let user 2 split its message into $s_{2,1}$ and $s_{2,2}$, and they are allocated to power $P_{2,1}$ and $P_{2,2}$, respectively, where $P_{2,1} + P_{2,2} = P_2$. In the splitting one user case, we know that $r_2 = \log_2\left(1 + \frac{P_{1,1}|h_1|^2 + P_{1,2}|h_1|^2 + P_2|h_2|^2}{\sigma^2}\right) - r_1$ because of the chain rule. If splitting two user can achieve a better 'point' (r'_1, r'_2) , this point should be at least $r'_1 > r_1$ when $r'_2 = r_2$, or $r'_2 > r_2$ when $r'_1 = r_1$. With the same transmission power,

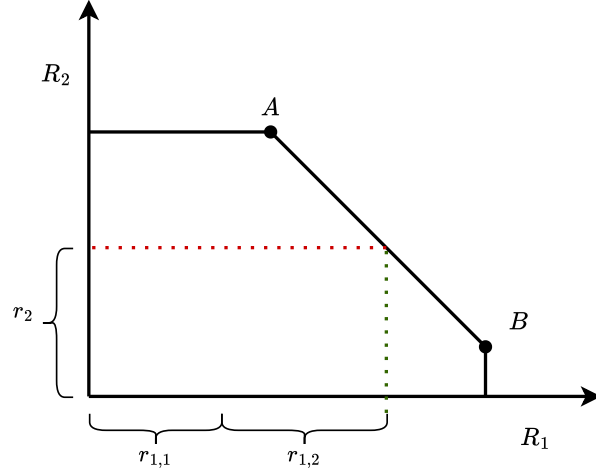


Fig. 2.1. An illustration of the two-user rate region of MAC.

the sum rate of r'_1 and r'_2 is $r'_1 + r'_2 = \log_2 \left(1 + \frac{P_{1,1}|h_1|^2 + P_{1,2}|h_1|^2 + P_2|h_2|^2}{\sigma^2} \right)$. If $r'_1 > r_1$, $r'_2 < \log_2 \left(1 + \frac{P_{1,1}|h_1|^2 + P_{1,2}|h_1|^2 + P_2|h_2|^2}{\sigma^2} \right) = r_2$, which contradicts the assumption. Similarly, it is impossible to find a (r'_1, r'_2) satisfying $r'_2 > r_2$ when $r'_1 = r_1$. Therefore, it is sufficient to split one user to achieve all the points of the rate region. The above analysis shows one of the main advantages of RSMA compared to OMA and NOMA. OMA cannot achieve all the points of the rate region because it only allocates dedicated resources to the users. Although NOMA allows users to share the resources, it can only achieve point A and point B in Fig. 2.1 without time-sharing.

Then, we extend this two-user case to K -user case, and the transceiver is shown in Fig. 2.2. For a K -user case, splitting $K - 1$ users into two parts enable all the points of the rate region to be achievable [36], and the detailed proof is given by [36] for interested readers. The intuition is that the rate region of K -user case is a K -dimensional space, so to achieve any point of this K -dimensional space, it is sufficient to have $K - 1$ degrees of freedom and splitting two parts are enough for being adjustable for each dimension. On the transmitting side, each user splits its messages and then encodes them into streams. Without loss of generality, we assume that each user splits its message into two parts. We note that actually splitting $K - 1$ users enables

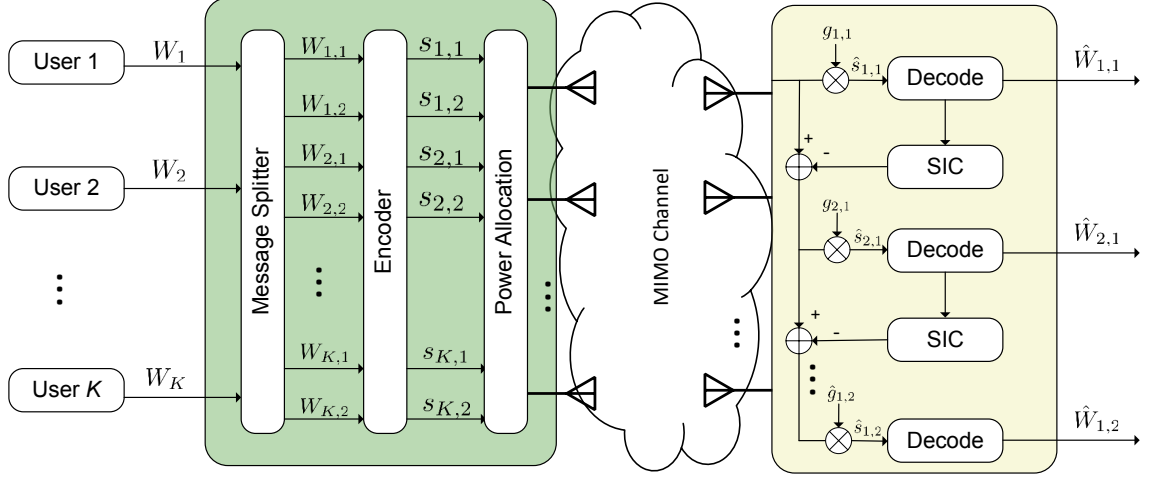


Fig. 2.2. Transceiver architecture of K -user uplink RSMA.

RSMA to achieve all the points of the rate region [36], and this is equivalent to one user allocating zero power to one of the split streams. On the receiving side, the BS uses SIC to decode the received streams. We use h_k to denote the channel between user k and BS, and $P_{k,i}$ is the power allocated to stream $s_{k,i}$. and the received signal at BS is

$$y = \sum_{k=1}^K \left(h_k \sum_{i=1}^2 \sqrt{P_{k,i}} s_{k,i} \right) + n. \quad (2.8)$$

To obtain the SINR of each received stream, firstly the decoding order needs to be decided. Let $\pi(k,i)$ denote the map function that maps the stream $s_{k,i}$ to its decoding order, i.e. $\pi(s_{k,i}) = j$ represents that $s_{k,i}$ is the j -th decoded stream at BS. For a stream of user k , $s_{k,i}$, $i \in \{1,2\}$, its SINR is

$$\rho_{k,i} = \frac{P_{k,i} |h_k|^2}{\sum_{\pi(m,i) > \pi(k,i)} P_{m,i} |h_m|^2 + \sigma^2}, \quad (2.9)$$

where the first term of the denominator is the summation of the interference from the streams which are decoded after $s_{k,i}$. The achievable rate of $s_{k,i}$ is

$$r_{k,i} = \log_2(1 + \rho_{k,i}), \quad (2.10)$$

and the achievable rate of user k is

$$r_k = \sum_{i=1}^2 r_{k,i}. \quad (2.11)$$

The advantages of uplink RSMA mainly come from the flexible decoding order and interference management. However, designing an appropriate decoding order and power allocation strategy are also challenges in uplink RSMA. Especially the decoding order, it is not involved in downlink RSMA which always decodes the common stream first and then decodes the private stream, and it drives us to investigate uplink RSMA. By designing the decoding order and power allocation, uplink RSMA can meet different goals, such as enhancing user fairness [21, 22, 38, 39] in MIMO, finite blocklength regime. Uplink RSMA also shows its advantages in improving spectrum efficiency [23, 37, 47]. [42] shows that uplink RSMA has the potential to be applied in low latency applications with low complexity.

2.2 Network Slicing and Models of eMBB, URLLC and mMTC

This subsection gives a brief introduction to network slicing and the models of eMBB, URLLC and mMTC, and then a literature review follows.

5G features three generic services: eMBB, URLLC, and mMTC. These three services have various applications and requirements [48, 49]. eMBB aims to provide extremely high data rates with high reliability, for applications such as augmented reality and remote presence. mMTC aims to provide connectivity for numerous devices which are cost and energy-constrained. The objective of URLLC is to provide a service featuring ultra-reliability and low latency while the rate is relatively low compared to mMTC. These three services have various applications and requirements [48, 49]. Some typ-

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

ical use cases are shown in Table 2.1 [50–57]. eMBB aims to provide service for users requiring high data rate, and it is suitable for uploading large files and AR/VR who require high throughput to ensure a smooth and immersive experience. URLLC is for applications with low latency and high reliability, such as factory automation involving robotic control and automated production lines, motion control in the autonomous driving. mMTC provides services for massive connectivity. For example, in smart factories, massive sensors are deployed for monitoring of temperature, humidity and machine status, and mMTC can be applied in such a scenario. Wearable devices are also typical mMTC applications because usually they require large-scale connectivity.

As introduced in Chapter 2.2, Fig. 2.3a is the illustration of OMA. Each service occupies individual resources without overlap, but users of the same service can share the resources, i.e. one URLLC user can be allocated an individual resource denoted by light blue, and multiple URLLC users can also be allocated to the same resources represented by dark blue. Fig. 2.3b shows the general non-orthogonal slicing scenario, which includes NOMA and RSMA. Different services can share the same resource, and the sharing part is denoted by the overlap.

Network slicing is a promising technique to satisfy different service requirements [2,3]. This technique allocates communication resources to the users requiring different

services to guarantee their varied requirements. Usually, network slicing is done by OMA-based slicing, which is shown in Fig. 2.3a. Each service has its dedicated time-frequency resource, but users of the same service can share the resources. For example, a URLLC user can occupy an individual resource denoted by light blue, or multiple URLLC users can share the same resource which is denoted by dark blue. However, as the number of users grows, a more flexible and efficient scheme may be explored, and these three services have different requirements, so network slicing may leverage these differences to obtain better performance. Therefore, NOMA-based slicing is investigated [43] as shown in Fig. 2.3b, where different services can share the same time-frequency resources. We use black boxes to denote the sharing resources. For example, the sharing resources between eMBB and mMTC are represented by the green part in the black box, and the sharing resources between eMBB and URLLC are represented by the blue parts in the black boxes.

Several works on NOMA-based network slicing scheme have shown the potential to improve the performance in the uplink, and they were studied in [43, 58, 59]. Applying NOMA-based slicing for eMBB and URLLC coexistence in downlink has also been studied and showed lower power consumption and more effective utilization [60–66]. We note that our discussion is limited to the wireless part of the network slicing, and sometimes it is referred to as spectrum slicing [67].

We consider that there are F frequency channels indexed by $f \in \{1, \dots, F\}$, and S time slots indexed by $s \in \{1, \dots, S\}$, as shown in Fig. 2.3. To meet the requirements of eMBB, URLLC and mMTC, we make the following assumptions. eMBB users are allocated to single or multiple radio resources at one or several given frequencies, while URLLC users can occupy several radio frequencies and a single mini-slot because of the stringent latency requirement¹, and mMTC users are allocated a specified fre-

¹Besides latency requirement, high reliability is also an important requirement of URLLC and diversity is critical to achieve it [68]. Therefore, transmitting with multiple frequencies would be more reasonable to increase diversity. In this way, increasing diversity compensates for fades in wireless channels [69, 70] and guarantees reliability.

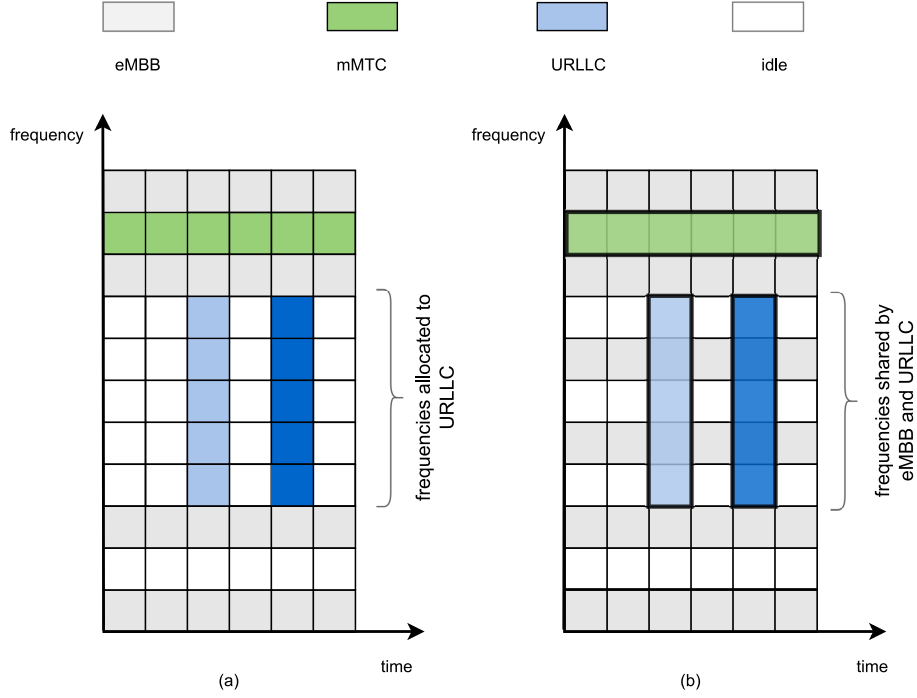


Fig. 2.3. Illustration of network slicing. (a) is the OMA scenario, and each service is allocated isolated resources, but the users of the same services can share the same time and frequency resources, i.e. one URLLC user can occupy an isolated resource denoted by light blue, and multiple URLLC can also share the same resource which is denoted by dark blue. (b) is the general non-orthogonal slicing scenario, i.e., NOMA and RSMA. mMTC and eMBB, eMBB and URLLC can share the resources, and the sharing parts are denoted with black boxes.

quency resource. The numbers of URLLC users and active mMTC devices are n_U and n_M , respectively. The arrival process of mMTC follows a Poisson distribution, $n_M \sim \text{Poisson}(\lambda_M)$, where λ_M is the mean value and referred to as the arrival rate of mMTC users.

We use the setup the same as in [43] and it is as follows. The radio access and competition among eMBB devices have been solved prior to the considered time slot, and there is only one eMBB device in each time slot. Each frequency channel f is assumed to be within the time-and frequency-coherence interval of the wireless channel, so the wireless channel coefficients are seen as constant. The channel coefficients of eMBB user, URLLC user u and mMTC user m at frequency f are denoted by $h_{B,f} \in \mathbb{C}$, $h_{U,u,f} \in \mathbb{C}$ and $h_{M,m,f} \in \mathbb{C}$, respectively, where $u = 1, 2, \dots, n_U$ and $m = 1, 2, \dots, n_M$, and they fade independently. The channels are considered as Rayleigh fading channels, i.e.

$h_{B,f} \sim \mathcal{CN}(0, \Gamma_B)$, $h_{U,u,f} \sim \mathcal{CN}(0, \Gamma_U)$, and $h_{M,m,f} \sim \mathcal{CN}(0, \Gamma_M)$, where Γ_B , Γ_U and Γ_M are the average channel gains for eMBB, URLLC and mMTC users, respectively. We mainly focus on long-term resource management and the differences in the path loss are accounted for through the average channel gains. The channel gains of the users of three services in radio frequency f are denoted by $G_{B,f} = |h_{B,f}|^2$, $G_{U,u,f} = |h_{U,u,f}|^2$, and $G_{M,m,f} = |h_{M,m,f}|^2$. Without loss of generality, the power of the noise at BS is normalized to one. Thus, the received power equals Signal-to-Noise Ratio (SNR). The received signal at BS at time slot s and frequency f can be represented as

$$y_{s,f} = h_{B,f}x_{B,s,f} + \sum_{u=1}^{n_U} h_{U,u,f}x_{U,u,s,f} + \sum_{m=1}^{n_M} h_{M,m,f}x_{M,m,s,f} + z_{s,f}, \quad (2.12)$$

where $x_{B,s,f}$, $x_{U,u,s,f}$, $x_{M,m,s,f}$ and $z_{s,f}$ are signals transmitted by eMBB user, URLLC user u , mMTC user m and noise at the frequency f and time slot s , respectively.

We assume that the eMBB device has perfect CSI, while URLLC and mMTC do not have CSI, and BS also has perfect CSI as in [43]. Thus, the eMBB device can adjust the transmission power according to the CSI, while URLLC and mMTC devices cannot adapt power because of lack of CSI and the transmission power is fixed and normalized to 1. The reliability requirements of eMBB, URLLC and mMTC are ϵ_B , ϵ_U and ϵ_M , respectively. Let E_B , E_U , E_M denote the events that the eMBB user is in outage, the event URLLC users are in outage and the event mMTC users are in outage, respectively. Thus, the outage probabilities should satisfy these reliability requirements, i.e. $\Pr(E_B) \leq \epsilon_B$, $\Pr(E_U) \leq \epsilon_U$ and $\Pr(E_M) \leq \epsilon_M$.

To investigate the slicing scheme, we introduce the models of eMBB, URLLC and mMTC individually to understand their different properties. In the following subsections, models of eMBB, URLLC and mMTC are introduced, and at this stage, they are considered individually without interference from other services.

2.2.1 eMBB

Since the eMBB device and BS are assumed to have perfect CSI, the eMBB device can adjust the transmission power $P_B(G_{B,f})$ according to the instantaneous channel gains to achieve a predefined SNR value. Thus, the objective is that the eMBB device transmits at the largest rate $r_{B,f}$ at frequency f and the outage probability is lower than reliability requirements ϵ_B under a long-term average power constraint, so the problem can be formulated as

$$\begin{aligned} \max \quad & r_{B,f} \\ \text{s.t.} \quad & \Pr[\log_2(1 + G_{B,f}P_B(G_{B,f})) < r_{B,f}] \leq \epsilon_B \\ & \mathbb{E}[P_B(G_{B,f})] = 1. \end{aligned} \tag{2.13}$$

The optimal solution to this problem is given by truncated power inversion [71], which means that the eMBB device chooses a transmission power that is inversely proportional to the channel gain $G_{B,f}$ if it is above a given threshold $G_{B,f}^{min}$. The probability that the eMBB transmits is

$$a_B = \Pr[G_{B,f} \geq G_{B,f}^{min}] = e^{-G_{B,f}^{min}/\Gamma_B}. \tag{2.14}$$

Without interference from other services, the only outage source is that the eMBB user does not transmit due to the insufficient SNR level, so the event that the eMBB user is in outage is actually the event that the eMBB user is inactive, i.e. not transmitting. Thus, the error probability is

$$\Pr(E_B) = 1 - a_B. \tag{2.15}$$

Impose the reliability requirement ϵ_B

$$\Pr(E_B) = \epsilon_B, \tag{2.16}$$

so we can obtain the value of threshold SNR

$$G_{B,f}^{min} = \Gamma_B \ln \left(\frac{1}{1 - \epsilon_B} \right). \quad (2.17)$$

With the power-inversion-scheme, the transmission power is chosen as a function of the instantaneous channel gain $G_{B,f}$ as

$$P_B(G_{B,f}) = \begin{cases} \frac{G_{B,f}^{tar}}{G_{B,f}} & \text{if } G_{B,f} \geq G_{B,f}^{min} \\ 0 & \text{if } G_{B,f} < G_{B,f}^{min} \end{cases}, \quad (2.18)$$

where $G_{B,f}^{tar}$ is the target SNR. Imposing the average power constraint to be normalized to one,

$$1 = \mathbb{E}[P_B(G_{B,f})] = \int_{G_{B,f}^{min}}^{\infty} \frac{1}{\Gamma_B} e^{-x/\Gamma_B} P_B(x) dx, \quad (2.19)$$

and the target SNR can be obtained as

$$G_{B,f}^{tar} = \frac{\Gamma_B}{\Gamma \left(0, \frac{G_{B,f}^{min}}{\Gamma_B} \right)}. \quad (2.20)$$

Thus, the largest transmission rate of eMBB can be obtained as

$$r_{B,f} = \log_2(1 + G_{B,f}^{tar}) \text{ (bits/s/Hz)}. \quad (2.21)$$

From this result, we know that $r_{B,f}$ only depends on Γ_B and ϵ_B , so for each frequency channel the largest transmission rate is the same, and the index f can be omitted. This largest transmission r_B is referred to as r_B^{orth} ,

$$r_B^{orth} = \log_2(1 + G_{B,f}^{tar}) \text{ (bits/s/Hz)}. \quad (2.22)$$

2.2.2 URLLC

URLLC devices transmit messages across F_U frequency resources, where $F_U \leq F$. We assume that URLLC devices do not have CSI, and the transmission power is fixed to 1. BS has the CSI of URLLC users and uses SIC to decode the messages. In the SIC procedure, BS receives the superposition of URLLC signals, successively detects the received data symbols, and then subtracts them from the superposition of signals. URLLC users are intermittent and each user is active with the probability a_U , so the number of active URLLC users is random in a given spanning time resource with multiple frequency resources, which we refer to as a frequency strip. This is a generalization of URLLC setup in [43], which assumes a single URLLC user. Here we consider the worst case that there are always n_U URLLC users transmitting simultaneously in the given frequency strip, so the outage probability in practice should be lower than the estimation which will be given then. Let $G_{U,u,f}$ denote the channel gain of user u in frequency channel f , and the user decoded after user u is denoted by $j > u$, so the SINR of user u in frequency f

$$\sigma_{U,u,f} = \frac{G_{U,u,f}}{1 + \sum_{j>u}^{n_U} G_{U,j,f}}. \quad (2.23)$$

Thus the outage probability is

$$\Pr(E_U) = \frac{1}{n_U} \sum_{u=1}^{n_U} \Pr\left(\frac{1}{F_U} \sum_{f=1}^{F_U} \log_2(1 + \sigma_{U,u,f}) < r_U\right), \quad (2.24)$$

where r_U is the target rate of URLLC.

The decoding order is decided by the sum of the mutual information across F_U frequency resources [58]. The users are successively decoded, and BS picks the device with the largest value of the sum of mutual information I_{sum} to decode at each step,

i.e., for user u it is

$$I_{sum} = \frac{1}{F_U} \sum_{f=1}^{F_U} \log_2 \left(1 + \frac{G_{U,u,f}}{1 + \sum_{j>u}^{n_U} G_{U,j,f}} \right). \quad (2.25)$$

Thus, we can impose the reliability requirement $\Pr(E_U) = \epsilon_U$ and obtain the maximum r_U , and the sum-rate of URLLC users is

$$r_U^{sum} = n_U r_U \text{ (bits/s/Hz)}. \quad (2.26)$$

Increasing F_U can enhance the frequency diversity, so it is possible to obtain a higher r_U while satisfying the reliability requirement ϵ_U with larger F_U .

2.2.3 mMTC

The arrival process of mMTC devices is random, so we assume that it follows a Poisson distribution with the arrival rate λ_M , and the rate of mMTC devices is fixed at r_M . CSI is unknown for mMTC, and the transmission power is fixed to 1. Since mMTC devices are allocated a specified frequency resource, the index f is omitted.

BS uses SIC to decode mMTC devices. The decoding order is decided by the channel gains of users. Let m denote the mMTC device with the m -th largest channel gain and the channel gains are sorted in descending manner $G_1 \geq G_2 \geq \dots G_{n_M}$, so the decoding order is $1, 2, \dots, n_M$. Since there is no interference from other services, for a random user m_0 , the SINR is

$$\sigma_{m_0} = \frac{G_{m_0}}{1 + \sum_{m=m_0+1}^{n_M} G_m}. \quad (2.27)$$

If $\log_2(1 + \sigma_{m_0}) \geq r_M$ holds, this mMTC device can be decoded successfully and then cancelled. Otherwise, it cannot be decoded and the procedure terminates. Let D_M

denote the number of successfully decoded devices, and the error probability measures the ratio of successfully decoded devices D_M to all active devices n_M . Thus the error probability is

$$\Pr(E_M) = 1 - \frac{\mathbb{E}[D_M]}{\lambda_M}, \quad (2.28)$$

where $\mathbb{E}[D_M]$ is the expectation of D_M , and λ_M can represent the mean value of n_M . Thus, the problem can be formulated as

$$\lambda_M^{orth}(r_M) = \max\{\lambda_M : \Pr(E_M) \leq \epsilon_M\}, \quad (2.29)$$

and the value of $\lambda_M^{orth}(r_M)$ can be computed by Monte Carlo method.

OMA-based slicing scheme is a conventional scheme, but it may not meet the requirements of future networks. NOMA-based network slicing scheme has the potential to improve the performance in the uplink, and it was studied in [43, 58, 59]. In [43], a NOMA-based slicing scheme was explored, and this work showed that NOMA-based slicing is more suitable than OMA-based slicing in some scenarios. In [58], network slicing for eMBB and URLLC coexistence was analysed. There was one eMBB user and multiple URLLC users in this scenario. When URLLC users had better channel conditions than the eMBB user, NOMA-based slicing always outperformed OMA-slicing. While when eMBB users had better channel conditions, NOMA-based slicing only had better performance when the eMBB rate was relatively high. In [59], the eMBB and mMTC coexistence scenario was analysed. There was one eMBB user and multiple mMTC users. This paper analysed the achievable pairs of eMBB rate and mMTC arrival rate. It showed NOMA-based slicing can achieve the pairs that OMA-based slicing cannot, and vice versa. In [72], the performance trade-offs between the eMBB rate and the URLLC latency were investigated in NOMA-based slicing with both puncturing and SIC methods. Applying NOMA-based slicing in downlink has also been studied and showed lower power consumption and more effective utilization.

In [60–66], eMBB and URLLC coexistence was studied. In [60], NOMA-based slicing was employed to minimize the transmission power and it had lower power consumption than OMA-based slicing. [61–64] showed that NOMA-based slicing can leverage resources more effectively, and it has promising performance when resource allocation and scheduling are designed properly, and [65] proposed a learning-based approach for resource allocation. In [66], a scheduling algorithm for eMBB and URLLC coexistence in MIMO NOMA systems was studied, and it brought a higher spectrum utilization than the OMA counterpart.

Although NOMA-based network slicing showed advantages, the interference during the SIC procedure may affect performance. As a more flexible MA technique, uplink RSMA manages the interference better by a more flexible decoding order. However, there are few works studying applying RSMA to network slicing. An RSMA-based slicing scheme for eMBB and URLLC coexistence was proposed in [73], and this work showed that RSMA could achieve a larger rate region than NOMA-based and OMA-based slicing when the power splitting factor is properly configured. However, this work only discussed eMBB-URLLC coexistence when URLLC users have better channel conditions. A more overall study of applying RSMA to network slicing has not been done.

2.3 Semantic communications and Its Metric

This section gives a brief introduction to semantic communications and a state-of-the-art semantic communications for text transmission model, namely DeepSC [74]. Then, an important parameter of semantic text transmission, namely sentence similarity [75], and a data regression method to approximate it are introduced.

Semantic communication has attracted attention from both academia and industry [4–7]. Different from conventional bit communication, semantic communication

aims to transmit the semantic meaning of the original messages. In this way, the source data is compressed dramatically while the semantic meaning remains the same [5]. Semantic communication is not a new concept, and it was first proposed in [76]. In [76], Claude Shannon categorized communication problems into three levels: (1) *How accurately can the communication symbols be transmitted?* (2) *How precisely do the transmitted symbols convey the desired meaning?* (3) *How effectively does the received meaning affect conduct in the desired way?* Level 1 is the technical problem, Level 2 is the semantic problem, and Level 3 is the effectiveness problem, which measures to what degree the received message successfully drives the desired action. Semantic communications focus on the second level. Instead of just transmitting bits and ignoring the meaning, it focuses on transmitting the desired meaning with symbols and how to precisely present the meaning with symbols. Therefore, the semantic data should be accuracy of meaning and efficiency, because the goal is to transmit core meaning without unnecessary information while ensuring the sender and receiver share the same understanding. Some theoretical works have already been done. [77] gave the concept of the amount of semantic information based on logical probability functions and tentatively defined semantic noise, and semantic noise includes interpretation errors, i.e. the receiver misinterprets the semantic content of a message, contextual misalignment, i.e. the sender and the receiver do not share the same contextual knowledge, and ambiguity and vagueness of message, i.e. a message contains ambiguous expressions. [78] proposed a model-theoretical framework for semantic data compression and reliable semantic communication. [79] proposed a framework of semantic communication, and it obtained the optimal transmission policies which minimized the end-to-end average semantic error based on the Bayesian game.

Although some theoretical works of semantic communication have been done, most of the works in the general area of communications focused on the implementation of the first level of communication. The second level, the semantic problem, was not fully investigated. Thanks to the development of Deep Learning (DL) based Natural

Language Processing (NLP), the implementation of semantic communication is being explored. [80] designed a joint source-channel coding of text based on DL for transmitting the meaning of text. [81] proposed a joint source and channel code for wireless image transmission. [82] designed a DL-based system to recover the speech signal in semantic communication. [74] proposed a DL-based semantic communication system for text transmission to maximize the system capacity and minimize semantic error. [83] proposed a semantic communication framework for a multi-user scenario, and it unified the structure of transmitters for different tasks, image retrieval, machine translation, and visual question answering. [84] developed a DL-based semantic communication system which can serve many different tasks with multiple modalities of data.

As a potential MA scheme, there are works investigating downlink RSMA for semantic communication. In [85], the BS encoded common knowledge into the common stream and encoded the information for each user into the private stream, and a problem is formulated to minimize the communication and computation energy consumption. Simulation results showed that RSMA has higher energy efficiency. [86] designed a framework of RSMA for semantic-aware networks, and shows that RSMA has a lower Age of Incorrect Information (AoII) compared to SDMA. Although Age of Information (AoI) measures the freshness of information at the receiver, it fails to capture the information content of the updates and may lead to unnecessary penalties. For example, a message could be fresh but inaccurate because of the noise during the transmission, so AoI is small but the receiver does not receive the desired message. In addition, inaccurate information can lead the system to make wrong decisions if it is not corrected timely, but AoI cannot measure the duration of the incorrect information and the negative impact it brings. AoII is proposed to address the shortcomings of AoI by capturing the accuracy of information and timeliness by measuring the duration of incorrect information. [87,88] investigated RSMA for semantic communication based on probability graphs. Since compression brings additional computational resource consumption, [87] aimed to minimize the total consumption and [88] maximized the

sum of semantic rates subject to the transmission power, semantic compress ratio and rate allocation constraints. The results show that RSMA is energy efficient and can achieve a higher rate after compression compared to SDMA. [89] applied RSMA to a task-oriented semantic information transmission framework for image transmission and showed RSMA improved the quality of experience.

Semantic has shown its potential for future communication applications, but this does not mean that conventional bit communication will disappear in future networks. A prerequisite of applying semantic communication is that users only care about the meaning of the messages and do not care whether the received messages are the same as the transmitted ones, i.e. some sensors prefer transmitting semantic messages such as significant status changes rather than raw data. However, some users will still require original messages, i.e. some human users collect raw data for detailed analysis. Therefore, future networks should accommodate both bit users and semantic users. Although MA techniques have been studied widely in the bit communication regime, the design of MA techniques for semantic-user-and-bit-user coexistence scenarios is still in its infancy. This motivates us to investigate suitable MA schemes for the semantic-user-and-bit-user coexistence scenario. [90] studied the designs of NOMA in the downlink, [75, 91] studied the NOMA designs for the coexistence scenario in the downlink. [92] studied the coexistence of one bit user and one semantic user in the uplink.

This literature motivates us to explore that as an efficient MA technique, whether RSMA shows any benefit in the semantic-user-and-bit-user coexistence scenario. For the semantic communications part, we focus on the semantic communications framework for text transmission. In the next subsections, the architecture of DeepSC, the related metric and its approximation are introduced.

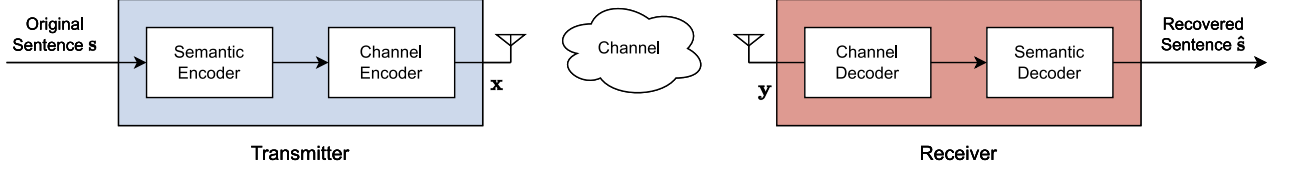


Fig. 2.4. The illustration of the framework of DeepSC.

2.3.1 Semantic Communication Metric

DeepSC is a state-of-the-art semantic text transmission framework, and its architecture is shown in Fig. 2.4. The transmitter of DeepSC has two parts, a semantic encoder and a channel encoder. The semantic encoder contains multiple Transformer [93] encoder layers and the channel encoder uses dense layers. First, the transmitter extracts the features of the original text $\mathbf{s} = [w_1, w_2, \dots, w_L]$, where w_l is the l -th word and L is the average number of words per sentence. The original text $\mathbf{s}$ is tokenized and converted to a word embedding vector and each element is a high-dimension vector. The semantic encoder extracts features with self-attention mechanism by capturing both local and global dependencies within the text, and encodes them into high-dimension semantic vectors [93]. The channel encoder maps the output of the semantic encoder to obtain the transmitted symbols and each symbol is in complex form. Since neural networks cannot process complex numbers, the channel encoder outputs real and imaginary parts separately, rather than a complex number directly. Therefore, a word is represented by several semantic symbols, and the transmitted signal is $\mathbf{x} \in \mathbb{C}^{1 \times KL}$, where K is the average number of semantic symbols per word. Considering that both the transmitter and receiver have one antenna, the received signal $\mathbf{y} = h\mathbf{x} + \mathbf{n}$, where $h \in \mathbb{C}^{1 \times 1}$ is the channel coefficient and $\mathbf{n} \in \mathbb{C}^{1 \times KL}$ is the noise. Then the receiver uses channel decoder and semantic decoder to recover the original sentence and the recovered sentence is denoted as $\hat{\mathbf{s}}$. Both the transmitter and the receiver are realized by neural networks, and they are jointly trained to make the meaning of $\hat{\mathbf{s}}$ the same as the meaning of the original sentence $\mathbf{s}$. Thus, a param-

eter that measures the similarity between the transmitted sentence and the received sentence is important for the system design, and DeepSC uses sentence similarity to measure the similarity. In [74], the sentence similarity is defined as

$$\xi(\hat{\mathbf{s}}, \mathbf{s}) = \frac{B_{\Phi}(\mathbf{s}) \cdot B_{\Phi}(\hat{\mathbf{s}})^T}{\|B_{\Phi}(\mathbf{s})\| \|B_{\Phi}(\hat{\mathbf{s}})^T\|}, \quad (2.30)$$

where B_{Φ} is Bidirectional Encoder Representations from Transformers, which is a NLP framework that helps computers understand the meaning of language proposed in [94]. $\xi(\hat{\mathbf{s}}, \mathbf{s})$ is a value between 0 and 1, and higher $\xi(\hat{\mathbf{s}}, \mathbf{s})$ represents higher sentence similarity. The sentence similarity is related to the architecture of the neural networks. For DeepSC, it is decided by SNR ρ of the received signal and K [74]. Thus, $\xi(\hat{\mathbf{s}}, \mathbf{s})$ can be represented by a function of ρ and K , i.e. $\xi(\hat{\mathbf{s}}, \mathbf{s}) = \varepsilon(K, \rho)$.

To further investigate the performance of the semantic communication network in [74], the semantic rate is proposed in [95]. Let I denote the average amount of information in the transmitted sentence $\mathbf{s}$, which is measured by the semantic unit (*sut*), so the information contained in each symbol is $\frac{I}{KL}$ (*suts/symbol*). Since the symbol rate is equal to the transmission bandwidth, the amount of transmitted information is $\frac{WI}{KL}$ (*suts/s*), where W is the transmission bandwidth. At the receiver side, the received amount of information is

$$S = \frac{WI}{KL} \varepsilon(K, \rho), \quad (2.31)$$

and S can describe the semantic rate (*suts/s*). I/L is related to the source type, and it is a constant value for a given type of source [95]. The value of $\varepsilon(K, \rho)$ is related to the design of neural network and a well-designed neural network would have a high $\varepsilon(K, \rho)$, and we can obtain it with a given K and ρ for DeepSC [74]. Part of the $\varepsilon(K, \rho)$ are presented in Fig. 2.5 with blue markers.

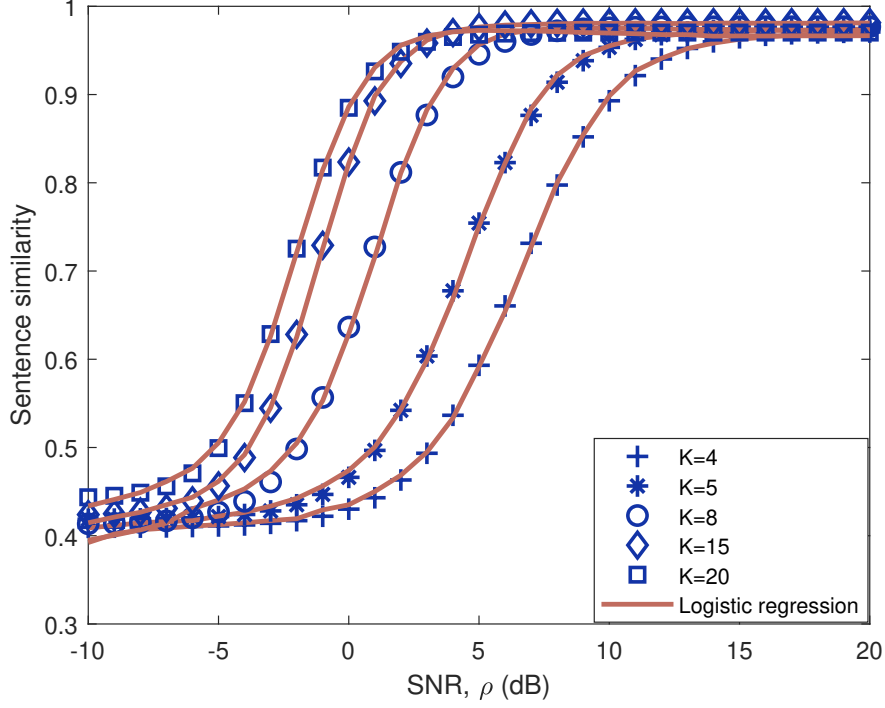


Fig. 2.5. The logistic regression for approximating $\varepsilon(\rho, K)$,

2.3.2 Data Regression Method For Semantic Rate Approximation

From the above subsection, we know how to obtain sentence similarity $\varepsilon(K, \rho)$ and semantic rate S , and how they relate to the average number of semantic symbols per word, K , and the received SNR, ρ . However, these relations are not presented mathematically, and this could be a problem for semantic communication design. Thus, [75] proposed a data regression method to approximate $\varepsilon(K, \rho)$.

From the blue marks in Fig. 2.5, we can observe that for a given K , $\varepsilon(K, \rho)$ is monotonically non-decreasing as ρ increases, and the derivative of $\varepsilon(K, \rho)$ with respect to ρ , $\frac{d}{d\rho}\varepsilon(K, \rho)$, increases first and then decreases. According to the observation, [75] used the generalized logistic function to approximate $\varepsilon(K, \rho)$, which is

$$\varepsilon(K, \rho) \approx \tilde{\varepsilon}_K(\rho) \triangleq A_{k,1} + \frac{A_{k,2} - A_{k,1}}{1 + e^{-(C_{k,1}\rho + C_{k,2})}}. \quad (2.32)$$

For a given K , $A_{k,1}$ and $A_{k,2}$ are the lower asymptote and the upper asymptote, respectively, and $C_{k,1} > 0$ denotes the logistic growth rate and $C_{k,2}$ decides the logistic mid-point. Then [75] applied the minimum mean square error criterion for fitting this generalized logistic function to get the parameters. Part of the results are shown in Fig. 2.5 with red lines, and they can approximate $\varepsilon(K, \rho)$ accurately. This approximation (2.32) will help us to design the semantic users and the bit user coexistence schemes. Since $\varepsilon(K, \rho)$ is only related to ρ for a given K , we omit K in the following sections for convenience, and our discussion is based on a specific K .

Although semantic communication shows its potential, bit communication will still play a role in future networks. This motivates us to investigate suitable MA for the semantic-user-and-bit-user coexistence scenario. As a promising MA technique, RSMA is a potential candidate, but RSMA for this coexistence scenario is still an open problem.

2.4 Hybrid-ARQ

HARQ is a critical mechanism to enhance reliability, increase diversity and improve the efficiency of packet-based transmission [12]. Thus, it is a crucial technique to meet the expectations in future networks [96, 97]. This mechanism is based on FEC and ARQ [13]. Generally, there are three types of HARQ: Type I HARQ, HARQ with Chase Combining (CC), and HARQ with incremental redundancy (IR). Type I HARQ uses the packet in the current round to decode, while HARQ with CC and HARQ with IR can decode packets in different rounds jointly. In HARQ with CC, if the receiver detects errors when it receives a packet, it sends a Negative Acknowledgment (NACK) to the transmitter, and then the transmitter sends an identical packet. At the receiver side, the receiver uses maximal ratio combining (MRC) to combine messages in every turn. In HARQ with IR, the transmitter initially sends a

packet with some amount of FEC parity bits. If the receiver detects errors and sends a NACK, the transmitter does not retransmit the same packet. Instead, it sends additional parity bits, which are systematically derived from the original data. The receiver combines all the received parity bits to reconstruct the original data [98]. CC is easier to implement while IR is more efficient.

HARQ has been extensively studied with both OMA [98,99] and NOMA [97,100–105]. For OMA, [98] optimized the transmission power for a given outage probability with limited retransmission times for the CC scheme, and [99] maximized the throughput in the IR scheme. For NOMA, [100–103] focused on the downlink scenario and showed that NOMA with HARQ can improve the outage probability and energy efficiency. [97, 104, 105] studied NOMA with HARQ in uplink, and showed that it can reduce the latency and save power consumption. Specifically, [97, 104, 105] showed NOMA with HARQ has promising performances for short packets, which means it is suitable for services requiring low latency. [106] introduces a backhauling approach employing NOMA with ARQ for enhancing the reliability of transmission and incorporates a packet repair technique to address the potential packet loss, and the results show that packet drop rate decreases dramatically and it ensures the reliability of the transmission. Performing HARQ orthogonally would not be able to accommodate numerous users with limited resources. It also increases the latency, because unaccommodated users need to wait in the queue. Clearly, allocating resources and performing HARQ non-orthogonally can improve the efficiency of systems and reduce the waiting time. Besides, BS can keep the previous unsuccessful copies of messages, so using a full payload to retransmit messages again would be wasteful. Meanwhile, fixed-size resources are preferred, so it is not practical to adjust the size of retransmission [97].

The literature studied that the design of HARQ with OMA and NOMA and shows non-orthogonal MA is more efficient when the number of users increases. Motivated by the literature, we investigate RSMA with HARQ to meet the high reliability re-

quirements of future networks.

Chapter 3

RSMA Facilitating Network Slicing

This chapter analyses the performances of different network slicing schemes in uplink. Network slicing aims to handle the services with different requirements [49, 107], and in the meanwhile it should ensure the performances of each services, so it may need to isolate services. The OMA-based network slicing scheme isolates different services completely, but it may not be efficient enough when the number of users increases. The NOMA-based scheme can let the users with different services share the same time frequency resources. For example, in Fig. 2.3, the three services, eMBB, URLLC and mMTC, can be performed in the same time slot, but they may be not at the same frequency. However, allocating three services to the same time frequency resources would violate the purpose of network slicing, because this is equivalent to not performing network slicing and it may bring negative impacts on performances. Therefore, only two services coexistence scenarios are considered in this chapter, eMBB and URLLC coexistence and eMBB and mMTC coexistence. URLLC and mMTC coexistence is not considered because mMTC is not sensitive to latency and URLLC can be scheduled before mMTC. Besides OMA-based and NOMA-based schemes, a promising scheme based on RSMA is studied. For eMBB and URLLC coexistence, the achievable rate pairs of eMBB and URLLC users are characterized. For eMBB

and mMTC coexistence, the achievable pairs of the eMBB rate and the arrival rate of mMTC are characterized. Simulation results show that for network slicing RSMA can outperform NOMA counterpart, and obtain significant gains over OMA in some regions. Detailed analysis are provided to explain the suitable scenarios for RSMA and why RSMA outperforms in those scenarios. Then this chapter concludes that RSMA is an efficient scheme for network slicing.

3.1 Introduction

Motivated by the increasing demands of users and the good properties of RSMA, this chapter studies the RSMA-based slicing¹ for network slicing. In RSMA scheme design, how the power is split between the two streams is a critical aspect because it dictates the performance. Splitting power fraction is used to refer to the power fraction allocated to the first stream without loss of generality. For eMBB and URLLC coexistence, the contributions of this paper are summarized as follows:

- This chapter investigates RSMA schemes for eMBB and URLLC coexistence and analyses the performances in different channel conditions. The achievable rate regions and the relation between the splitting power fraction and the sum-rate of URLLC users are shown. Although [73] analysed the eMBB and URLLC coexistence when channel conditions of URLLC users were better than eMBB, and showed that RSMA always outperformed OMA-based slicing² and NOMA-based slicing³ when users of the same services shared the resource in a non-orthogonal fashion, it did not analyse the opposite scenario when an eMBB user experiences a better channel than URLLC users. In the previous scenario, it would be easier for RSMA to exploit flexible decoding order; while

¹RSMA-based slicing is called RSMA for simplicity in the rest of this chapter.

²OMA-based slicing is called OMA for simplicity in the rest of the chapter.

³NOMA-based slicing is called NOMA for simplicity in the rest of the chapter.

in the later one, sometimes RSMA may boil down to NOMA. Interestingly, the later scenario shows when RSMA can outperform OMA and reveals a suitable scenario for RSMA. Numerical results show that RSMA can always achieve a larger rate region compared to NOMA, while it outperforms OMA in some scenarios. Adjusting splitting power fraction appropriately can improve the performance, and the lower the interference from the eMBB user, the higher the improvements RSMA can make.

- This chapter gives a detailed analysis of the suitable scenario for RSMA. The main superiority of RSMA in uplink comes from the flexible decoding order since the split streams do not have to be decoded successively. Hence, for the user with the lower rate, one of its split streams can be decoded after the streams of other users, so it has lower interference and achieves a higher rate; for the user with a higher rate, its stream can be decoded early to deal with more interference. Thus, all the users can achieve a higher rate. In this way, a higher decoding order flexibility leads to higher achievable rate pairs. This is the reason why RSMA has better performance when eMBB interference is not that high. The achievable rate regions of NOMA and RSMA in eMBB and URLLC coexistence are also simulated to demonstrate the suitable scenario of RSMA.

For eMBB and mMTC coexistence, the contributions are summarised below:

- This chapter proposes a RSMA scheme for eMBB and mMTC coexistence, and to the best of our knowledge, RSMA has not been applied to this scenario before. In this scheme, the message of the eMBB user is split into two streams. By adjusting the splitting power fraction of eMBB and scheduling the decoding order, the SINR of mMTC devices is also adjusted, since the first stream of eMBB message can be cancelled before decoding some mMTC devices. Consequently, more mMTC devices can have sufficient SINR to satisfy the target

rate compared to NOMA, which means a higher arrival rate of mMTC can be supported using RSMA for a given reliability requirement.

- This chapter analyses the performance of eMBB and mMTC coexistence with RSMA. The achievable pairs of the eMBB rate and the arrival rate of mMTC and the relation between the splitting power fraction of eMBB and the arrival rate of mMTC are shown. According to the simulation results, RSMA can achieve rate pairs non-achievable by NOMA and OMA, and RSMA is shown to leverage the difference between reliability requirements of different services better compared to NOMA.

3.2 System Model

This section introduces a general signal model of RSMA in the network slicing scenario. Different services can share the same radio resources when they communicate to a common BS as shown in Fig. 3.1. As introduced in Chapter 2.2, Fig. 2.3a is the illustration of OMA. Each service occupies individual resources without overlap, but users of the same service can share the resources, i.e. one URLLC user can be allocated an individual resource denoted by light blue, and multiple URLLC users can also be allocated to the same resources represented by dark blue. Fig. 2.3b shows the general non-orthogonal slicing scenario, which includes NOMA and RSMA. Different services can share the same resource, and the sharing part is denoted by the overlap.

The main idea of rate-splitting is to divide a message of one user into two parts, which is equivalent to adding a virtual user [36]. Thus, it provides a more flexible decoding order. For a general RSMA network slicing model in uplink, we assume that there is one eMBB user, two URLLC users and n_M mMTC users. For the users of the same services, they can share the same time and frequency resources and they are decoded by SIC at the receiver side. This setup can effectively capture the essence of the

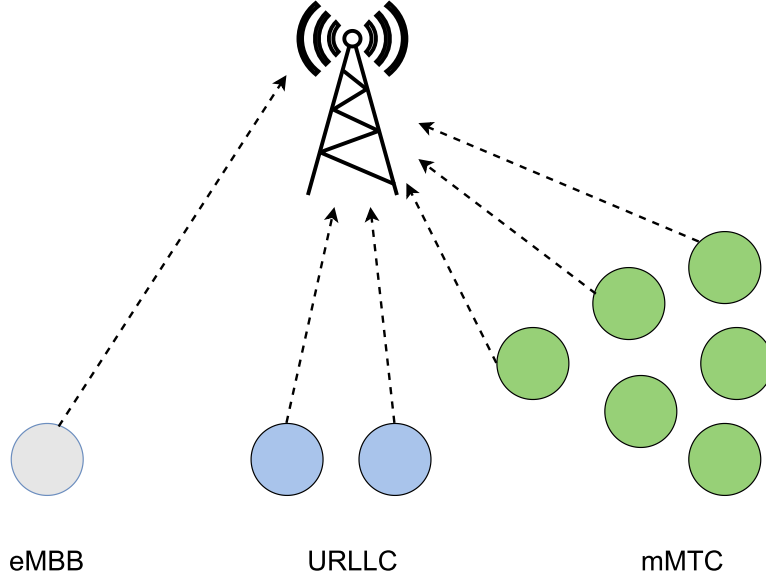


Fig. 3.1. The considered scenario that devices from three services transmit to a common BS in the uplink.

problem, because in a given cell the number of connected mMTC devices can be very large compared to the number of connected eMBB and URLLC users in each time slot. Since the arrival process of mMTC is random and mMTC devices are energy and cost-constrained, we assume that mMTC does not split the messages and the transmission power of the signal is fixed to 1. Hence, eMBB and URLLC users can have their message split, in contrast to mMTC users. Thus, the received signal of BS at a given time slot s and frequency f is (s and f are omitted for simplicity)

$$y = h_B \sum_{i=1}^2 \sqrt{P_{B,i}} x_{B,i} + \sum_{u=1}^2 \sum_{i=1}^2 h_{U,u} \sqrt{P_{U,u,i}} x_{U,u,i} + \sum_{m=1}^{n_M} h_{M,m} x_{M,m} + z, \quad (3.1)$$

where $x_{B,i}$ and $x_{U,u,i}$ are split messages of eMBB user and URLLC user u , respectively, and $P_{B,i}$, $P_{U,u,i}$ are the power allocated for $x_{B,i}$ and $x_{U,u,i}$, respectively. We emphasize that (3.1) is a general expression of the received signal at a given time and frequency grid, and this does not mean that all three services will share this time and frequency grid. If no eMBB user is scheduled at this given time and frequency grid, $x_{B,i}, i \in \{1, 2\}$ will be zero. Similarly, if URLLC and mMTC users are not scheduled at this time and frequency grid, $x_{U,u,i}, u \in \{1, 2\}, i \in \{1, 2\}$ and $x_{M,m}, m \in \{1, 2, \dots, n_M\}$ will be zero,

respectively. For eMBB and URLLC, the power constraints are $\mathbb{E}[\sum_{i=1}^2 P_{B,i}] = 1$ and $\sum_{i=1}^2 P_{U,u,i} = 1$, respectively, where $\mathbb{E}[\cdot]$ is the expectation operator. Here we consider that BS and users have one antenna, and adding more antennas would help scale up this scheme further.

We note that RSMA can unify NOMA and OMA in a single MA framework. When the transmission power is all allocated to one of the split streams and the other stream is not allocated power, RSMA boils down to NOMA; while when all the transmission power is turned off except one service, RSMA becomes OMA. Thus, OMA and NOMA can be contained in RSMA if full power control is applicable for all the users. However, since URLLC and mMTC do not have the CSI and the transmission power is fixed, they do not have the full power control, so for network slicing, OMA is not a subset of RSMA.

Although RSMA increases the number of messages, the decoding order is more flexible and it can achieve any boundary points of the rate region as shown in [36]. Other methods that achieve the same region are NOMA with time-sharing and joint encoding/decoding [12]. However, implementing time-sharing requires multiple time slots [23] and induces communication overhead [12]. A joint encoding/decoding approach is not practical to implement due to the high decoding complexity of random codes [12, 36]. Since URLLC has stringent latency requirements and mMTC is aimed at cost and power-constrained devices, NOMA with time-sharing and joint encoding/decoding can be problematic in network slicing, so they will not be considered in the following discussions.

3.3 Slicing for eMBB and URLLC

In this section, the coexistence of eMBB and URLLC will be discussed, and mMTC traffic will not be considered here. For eMBB we focus on the standard scheduled

transmission phase, so we assume the radio access and competition among eMBB devices have been solved prior to the considered time slot, and there is only one eMBB device and two URLLC devices in each time slot.

3.3.1 OMA for eMBB and URLLC

In the OMA scenario, we assume that the eMBB device and URLLC devices are allocated individual frequency channels. Let F denote the total number of the frequency channels, and URLLC devices are allocated F_U frequency channels.

The performance of this system is specified in terms of (r_B^{sum}, r_U^{sum}) of eMBB sum-rate r_B^{sum} and sum-rate of URLLC r_U^{sum} while reliability requirements (ϵ_B, ϵ_U) are satisfied. The sum-rate of eMBB is obtained as

$$r_B^{sum} = (F - F_U) r_B^{orth} \text{ (bits/s/Hz)}, \quad (3.2)$$

where r_B^{orth} is obtained in (2.22). The rate of URLLC devices can be computed from (2.24) while satisfying $\Pr(E_U) \leq \epsilon_U$, and the sum-rate can be computed from (2.26).

3.3.2 NOMA for eMBB and URLLC

In the NOMA scenario, eMBB and URLLC devices share all F frequency channels. Here BS uses SIC to decode eMBB and URLLC users. Besides SIC, BS can adopt another decoding approach, punctured method approach [63]. This is a preemptive mechanism that once BS detects the presence of URLLC transmission, it will 'puncture' the ongoing eMBB transmission. Unlike TDMA, puncturing is not predefined and occurs dynamically. It is shown in [43] that the SIC method outperforms the puncturing method, so in this chapter we focus on applying SIC at the BS. Like the scenario mentioned in Section 2.2.2, the worst case that there is always URLLC traffic when

eMBB is transmitting is considered. The decoding order is set as URLLC devices are decoded before eMBB devices and the reasons are following. (1) URLLC devices are sensitive to latency. (2) The reliability requirement of URLLC is much more stringent than eMBB. If decoding eMBB first and this eMBB device fails, decoding URLLC users will fail. Thus, decoding eMBB first will require eMBB to satisfy the same reliability requirement as URLLC, and it is actually not possible for eMBB to satisfy it, so decoding eMBB first can deteriorate the reliability of URLLC. Hence, the decoding priority should be given to URLLC because it can leverage the reliability requirements difference. (3) eMBB aims to provide the service with a high data rate, so decoding eMBB after URLLC will let eMBB users experience less interference and achieve a higher rate. Therefore, the procedure is that URLLC devices are always decoded first. If it is successful, then an eMBB device will be decoded. Otherwise, the procedure terminates, because the interference from URLLC devices cannot be cancelled.

Since there is always interference from the eMBB device to URLLC devices, it would be better to choose a smaller eMBB target SNR to get better overall performance, i.e.

$$G_{B,f}^{tar} \leq \frac{\Gamma_B}{\Gamma\left(0, \frac{G_{B,f}^{min}}{\Gamma_B}\right)}, \quad (3.3)$$

so the achievable rate for eMBB is

$$r_{B,f} = \log_2(1 + G_{B,f}^{tar}) \text{ (bits/s/Hz)}, \quad (3.4)$$

and the sum-rate is

$$r_B^{sum} = \sum_{f=1}^F r_{B,f}. \quad (3.5)$$

Given the rate of eMBB $r_{B,f}$, the objective is to find the highest achievable rate r_U while reliability requirements are satisfied. The error probabilities of eMBB and URLLC are calculated individually. For eMBB, the error probability can be com-

puted by the Monte Carlo method, which is

$$\Pr(E_B) = 1 - \mathbb{E}[D_B], \quad (3.6)$$

where D_B is the number of successfully decoded eMBB devices.

Then the error probability of URLLC is formulated. The decoding order of URLLC can be decided by the same method as in Section 2.2.2, and assuming it is $1, 2, \dots, n_U$, so the SINR of the URLLC device is

$$\sigma_{U,u,f} = \frac{G_{U,u,f}}{1 + \sum_{j>u}^{n_U} G_{U,j,f} + \delta_f G_{B,f}^{tar}}, \quad (3.7)$$

where δ_f is an independent Bernoulli random variable with parameter a_B which is the active probability of the eMBB device. Since the reliability requirement of the eMBB user is also stringent, i.e. 10^{-3} , it means that the probability that the eMBB device is inactive is extremely low. Thus, we consider the most stringent scenario that eMBB is always active, which means $\delta_f = 1$ and there is always interference from the eMBB device. Thus, the error probability of URLLC can be presented by (2.24), where $\sigma_{U,u,f}$ is obtained from (3.7) and $F_U = F$. Then, impose ϵ_U , $\Pr(E_U) \leq \epsilon_U$, and r_U can be computed by the Monte Carlo method, and the sum-rate is $r_U^{sum} = n_U r_U$ (bits/s/Hz).

3.3.3 RSMA for eMBB and URLLC

Since URLLC devices are very sensitive to latency and have high reliability requirements, i.e. decoding of URLLC should not depend on decoding of other services, they should be decoded first, so rate-splitting is only used for URLLC devices⁴. Here we

⁴Actually, from an information theory perspective, all users but one (e.g. eMBB user and one of the two URLLC users, or the two URLLC users) split their messages to achieve the capacity region. However, in network slicing, latency and reliability should also be considered. Hence, due to similar reasons as in Section 3.3.2, URLLC will always be decoded before eMBB and RSMA is only applied to URLLC users.

consider the setup in Section 3.2 with one eMBB user and two URLLC users and eMBB is always decoded after URLLC. If any URLLC message cannot be decoded, the procedure terminates and the eMBB message is lost.

The eMBB part is the same as in Section 3.3.2. G_B^{tar} is chosen by the same method as in (3.3), and the rate can be obtained from (3.4) and (3.5). The objective is to find the highest achievable rate pairs (r_U^{sum}, r_B^{sum}) while satisfying the reliability requirements of eMBB and URLLC. The error probability of eMBB can be obtained from (3.6). The decoding order of URLLC is the same as the one in Section 3.3.2. Let URLLC user 1 denote the device decoded first, and URLLC user 2 denote the other ⁵. The message from URLLC user 1 is divided into two parts, and the power factor for the first part is β , $\beta \in [0, 1]$. The received signal at BS at frequency f is

$$y = h_{B,f}x_{B,f} + \sqrt{\beta}h_{U,1,f}x_{U,1,1,f} + \sqrt{1-\beta}h_{U,1,f}x_{U,1,2,f} + h_{U,2,f}x_{U,2,f} + z, \quad (3.8)$$

where $x_{U,1,1,f}$ and $x_{U,1,2,f}$ are the signals from the first and second part of the message of URLLC user 1, respectively. Without loss of generality, we assume that $x_{U,1,1,f}$ is decoded first. The decoding order is set to $x_{U,1,1,f}$, $x_{U,2,f}$ and $x_{U,1,2,f}$, since this order can obtain all the boundary points of rate region by adjusting β . Thus, the SINR of $x_{U,1,1,f}$ is

$$\sigma_{U,1,1,f} = \frac{\beta G_{U,1,f}}{1 + (1-\beta)G_{U,1,f} + G_{U,2,f} + G_B^{tar}}, \quad (3.9)$$

the SINR of $\mathbf{x}_{U,2,f}$ is

$$\sigma_{U,2,f} = \frac{G_{U,2,f}}{1 + (1-\beta)G_{U,1,f} + G_B^{tar}}, \quad (3.10)$$

and the SINR of $\mathbf{x}_{U,1,2}$ is

$$\sigma_{U,1,2,f} = \frac{(1-\beta)G_{U,1,f}}{1 + G_B^{tar}}. \quad (3.11)$$

⁵Here we work with two assumptions on URLLC: (1) the instantaneous SNR is unknown and (2) the status of other URLLC user is unknown. Thus, BS will decide which user splits the message. Before exchanging data, the connection between URLLC users and BS needs to be set up. During this setup process, BS can obtain the CSI of URLLC users, and then decide which user will split the message and send the information of splitting message to this user.

Then we get the corresponding rate

$$r_{U,1,1,f} = \log_2(1 + \sigma_{U,1,1,f}) \text{ (bits/s/Hz)}, \quad (3.12)$$

$$r_{U,1,2,f} = \log_2(1 + \sigma_{U,1,2,f}) \text{ (bits/s/Hz)}, \quad (3.13)$$

$$r_{U,2,f} = \log_2(1 + \sigma_{U,2,f}) \text{ (bits/s/Hz)}, \quad (3.14)$$

and the sum-rate of URLLC user 1 is

$$r_{U,1} = \frac{1}{F} \sum_{f=1}^F (r_{U,1,1,f} + r_{U,1,2,f}) \text{ (bits/s/Hz)}. \quad (3.15)$$

Similarly, the rate of URLLC user 2 is

$$r_{U,2} = \frac{1}{F} \sum_{f=1}^F r_{U,2,f} \text{ (bits/s/Hz)}. \quad (3.16)$$

The error probability is formulated as

$$\Pr(E_U) = \frac{1}{2} \sum_{u=1}^2 \Pr(r_{U,u} < r_U), \quad (3.17)$$

and impose ϵ_U , $\Pr(E_U) \leq \epsilon_U$, and r_U can be computed by Monte Carlo method, and r_U^{sum} can be obtained from (2.26).

Remark 1. *The scenario with two URLLC users can also give insight into a more general scenario with n_U URLLC users. First, we consider a generic two-URLLC-user case, and we do not consider NOMA with time-sharing for the reasons mentioned in 3.2. The two URLLC users are not aware of CSI, so they are not able to adapt power and their transmission power is fixed and normalized to 1. We assume that user 1 splits the message into two streams, and user 2 does not split. If BS decodes the split streams of user 1 not successively, this decoding order can achieve any point on the diagonal line of the rate region. Fig. 3.2 is an illustration of rate regions of TDMA,*

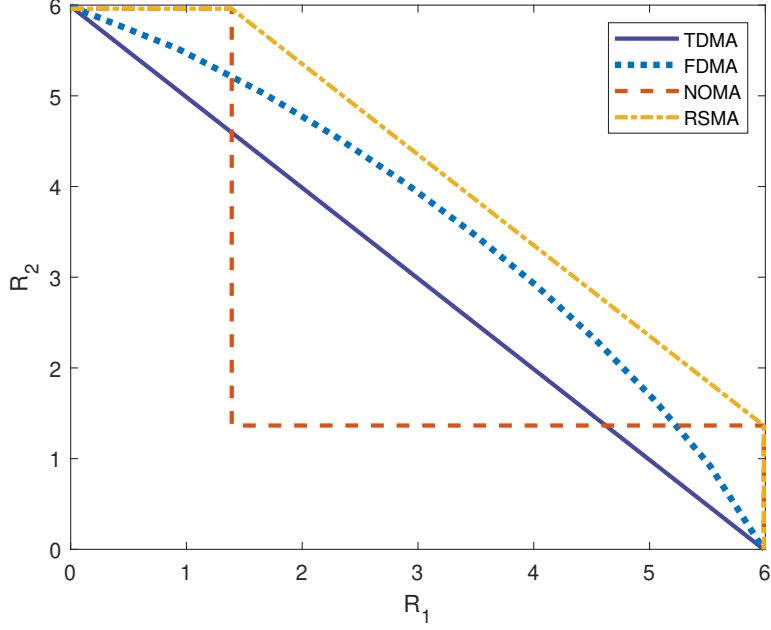


Fig. 3.2. An illustration of rate regions of TDMA, FDMA, NOMA and RSMA for a generic two-user case. The average channel gain of two users is 20 dB.

FDMA, NOMA and RSMA for a generic two-user case, and the average channel gain of both users is 20 dB. With the nonsuccessive decoding order, RSMA can achieve the points on the yellow diagonal line. Similarly, in URLLC and eMBB coexistence, RSMA could achieve all the boundary points of rate region by adjusting β . When there are two users, one user splits the message and these split streams are not decoded successively, this enables achieving any point on the diagonal line of the rate region by adjusting β . Thus, intuitively, when the number of users is n_U , these users can be ordered by mutual information in descending manner, and the order is $1, 2, \dots, n_U$. Let the first $n_U - 1$ users split messages and decode these split streams not successively, so this decoding order should also bring freedom to obtain any boundary points of the rate region, i.e. the decoding order could be $x_{U,1,1}, x_{U,2,1}, \dots, x_{U,n_U-1,1}, x_{U,n_U}, x_{U,n_U-1,2}, \dots, x_{U,2,2}, x_{U,1,2}$.

3.4 Slicing for eMBB and mMTC

In this section, the coexistence of eMBB and mMTC will be discussed. The arrival process of mMTC follows Poisson distribution as in Section 2.2.3, and URLLC traffic will not be considered. Similarly, there is one eMBB device and multiple mMTC devices, because we assume that the radio access and competition among eMBB devices are solved. Since only a single frequency channel is considered, the frequency indices f are omitted in this section.

3.4.1 OMA for eMBB and mMTC

In the OMA scenario, eMBB and mMTC users share the resources in a time-sharing manner. Let $\alpha, \alpha \in [0, 1]$, denote the time fraction allocated to eMBB users, and the time allocated to mMTC users is $1 - \alpha$. Thus, the achievable rate of eMBB is

$$r_B = \alpha r_B^{orth} \text{ (bits/s/Hz)}, \quad (3.18)$$

where r_B^{orth} is obtained in (2.22). The arrival rate λ_M can be computed by a method similar to that in Section II-C. It is equivalent to the scenario in Section II-C with mMTC rate at $\frac{r_M}{1-\alpha}$, so

$$\lambda_M = \lambda_M^{orth} \left(\frac{r_M}{1-\alpha} \right). \quad (3.19)$$

Then, λ_M can be obtained from (2.29).

3.4.2 NOMA for eMBB and mMTC

In the NOMA scenario, eMBB and mMTC devices share the same frequency resources at the same time. The decoding procedure is the same as in [43]. Trying to decode the eMBB message before mMTC messages may be an option since the

high rate requirement of eMBB can cause high interference when decoding mMTC messages. However, this is not the optimal option, because the error probability of mMTC is actually measuring the ratio of successfully decoded mMTC devices to all active mMTC devices instead of the decoding probability for each device. There may be some mMTC devices with very high channel gains, so it would be better to decode them before decoding the eMBB device.

Based on the analysis above, [43] gave the exact procedure. The mMTC devices are ordered in a descending manner, and BS tries to decode mMTC devices one by one. If an mMTC device cannot be decoded, BS tries to decode the eMBB device. If this eMBB device is successfully decoded, then BS continues decoding mMTC devices. Otherwise, the decoding procedure terminates, because eMBB interference cannot be cancelled.

For the eMBB device, similar to the eMBB and URLLC coexistence scenario, it would be better to choose a lower target SNR for eMBB G_B^{tar} which satisfies (3.3), so the rate of eMBB device is (3.4). Because the error of decoding the eMBB device is caused by the interference from mMTC devices and inactive state due to the insufficient SNR level, we must set a higher active probability for the eMBB device to satisfy the reliability requirement ϵ_B . For example, if there is not any other interference, like the scenario in Section 2.2.1, $a_B = 1 - \epsilon_B$, i.e. $\epsilon_B = 10^{-3}$, $a_B = 0.999$, while for eMBB and mMTC coexistence scenario, the active probability a_B should satisfy $a_B > 0.999$. Here we assume that $a_B = 1$, which means there is always eMBB interference for mMTC devices, and the error of eMBB is only caused by decoding error. Thus, the SINR of an mMTC device decoded before the eMBB device is

$$\sigma_{m_0} = \frac{G_{m_0}}{1 + \sum_{m=m_0+1}^{n_M} G_m + G_B^{tar}}. \quad (3.20)$$

If $\log_2(1 + \sigma_{m_0}) \geq r_M$, this mMTC device can be decoded. Otherwise, it failed and BS

will try to decode the eMBB device. The SINR of the eMBB device is

$$\sigma_B = \frac{G_B^{tar}}{1 + \sum_{m=m_0}^{n_M} G_m}. \quad (3.21)$$

If $\log_2(1 + \sigma_B) \geq r_B$, the eMBB device can be decoded, and then BS will try to decode other mMTC devices as in Section 2.2.3. Otherwise, it fails and the procedure terminates.

Let D_M and D_B denote the number of decoded mMTC devices and eMBB device, respectively. The error probability of mMTC is the ratio of expectation of decoded devices to all active devices, $1 - \mathbb{E}[D_M]/\lambda_M$, and the error probability of eMBB device is $1 - \mathbb{E}[D_B]$. In order to find the achievable pairs (r_B, λ_M) , the problem can be formulated as

$$\begin{aligned} \lambda_M^{\text{NOMA}}(r_B) = \max \{ & \lambda_M \geq 0 : \exists G_B^{tar} \\ \text{s.t. } & \mathbb{E}[D_M]/\lambda_M \geq 1 - \epsilon_M \\ & \text{and } \mathbb{E}[D_B] \geq 1 - \epsilon_B \}. \end{aligned} \quad (3.22)$$

For a given r_B , λ_M^{NOMA} can be computed by Monte Carlo method.

3.4.3 RSMA for eMBB and mMTC

According to the discussion in NOMA for eMBB and mMTC, the decoding procedure can terminate when neither mMTC device nor eMBB device can be decoded. This implies that mMTC device and eMBB do not have sufficient SINR to achieve the target rate, which is caused by the relatively high interference from each other. Then the decoding procedure terminates because it is 'lodged'.

To mitigate this effect, we apply RSMA in the eMBB and mMTC coexistence scenario. The eMBB message is split into two streams, so BS can decode one stream at a relatively low rate, and then continue decoding mMTC devices. Since part of

the interference from eMBB can be cancelled, the decoding procedure avoids getting ‘lodged’ to some extent. The whole decoding procedure is described below. The decoding order of the mMTC devices is defined according to the descending order of their channel gains. Firstly, BS tries to decode mMTC devices. If a mMTC device cannot be decoded, BS tries to decode the first stream of eMBB at a relatively low rate r_{B_1} , and then continues decoding mMTC devices. Similarly, if another mMTC device cannot be decoded, BS tries to decode the second stream of eMBB at rate r_{B_2} . If $r_{B_1} + r_{B_2} < r_B$, the eMBB device failed, and the decoding procedure terminates. Otherwise, BS continues decoding mMTC devices as in Section 2.2.3. Note that r_{B_1} is a non-negative value, so it is always possible to decode the first stream with a rate not higher than r_{B_1} , so we assume that BS could always decode the first stream regardless of whether the second stream can be decoded successfully. Similar to Section 3.4.2, the active probability of eMBB is $a_B = 1$, and the power factor for the first part of the eMBB message is β , $\beta \in [0, 1]$, a real number between 0 and 1. The received signal is

$$y = \sqrt{\beta} h_B x_{B_1} + \sqrt{1-\beta} h_B x_{B_2} + \sum_{m=1}^{n_M} h_m x_m + z, \quad (3.23)$$

where x_{B_1} and x_{B_2} are the signals from the first and second streams of the eMBB message, respectively, and h_m is the channel coefficient of mMTC user m . The target SNR of eMBB is still G_B^{tar} , and it satisfies (3.3). We assume that x_{B_1} is always decoded first, and its SINR at BS is

$$\sigma_{B_1} = \frac{\beta G_B^{tar}}{1 + (1-\beta) G_B^{tar} + \sum_{m=m_1}^{n_M} G_m}, \quad (3.24)$$

where m_1 represents the index of the mMTC device which will be decoded after x_{B_1} , and the rate is

$$r_{B_1} = \log_2(1 + \sigma_{B_1}) \text{ (bits/s/Hz)}. \quad (3.25)$$

The SINR of x_{B_2} is

$$\sigma_{B_2} = \frac{(1-\beta)G_B^{tar}}{1 + \sum_{m=m_2}^{n_M} G_m}, \quad (3.26)$$

where m_2 represents the number of the mMTC device which will be decoded after x_{B_2} , and the rate is

$$r_{B_2} = \log_2(1 + \sigma_{B_2}) \text{ (bits/s/Hz)}. \quad (3.27)$$

For mMTC devices decoded before all eMBB streams, the SINR is

$$\sigma_{m_0} = \frac{G_{m_0}}{1 + G_B^{tar} + \sum_{m=m_0+1}^{n_M} G_m}. \quad (3.28)$$

For mMTC devices decoded between x_{B_1} and x_{B_2} , the SINR is

$$\sigma_{m_0} = \frac{G_{m_0}}{1 + (1-\beta)G_B^{tar} + \sum_{m=m_0+1}^{n_M} G_m}. \quad (3.29)$$

For mMTC devices decoded after all eMBB streams, the SINR is as same as (2.27).

Let D_M and D_B denote the number of decoded mMTC devices and eMBB device, respectively. In this scenario, $D_B = 1$ holds only when $r_{B_1} + r_{B_2} \geq r_B$ holds. Otherwise, $D_B = 0$. For mMTC devices, they can be decoded when $\log_2(1 + \sigma_{m_0}) \geq r_M$ holds. Similarly, in order to find the achievable pairs (r_B, λ_M) , the problem can be formulated as

$$\begin{aligned} \lambda_M^{\text{RSMA}}(r_B) = \max \{ & \lambda_M \geq 0 : \exists G_B^{tar} \text{ and } \beta \\ & \text{s.t. } \mathbb{E}[D_M]/\lambda_M \geq 1 - \epsilon_M \\ & \text{and } \mathbb{E}[D_B] \geq 1 - \epsilon_B \}. \end{aligned} \quad (3.30)$$

λ_M^{RSMA} can be computed by the Monte Carlo method.

3.5 Numerical Results

This section presents the simulation results of eMBB and URLLC coexistence and eMBB and mMTC coexistence. Since we focus on the long-term resource management and the differences in the path loss are accounted for through the average channel gains, here we use average channel gains to represent the strengths of channels for concision. For eMBB and URLLC coexistence, the achievable rate pairs of eMBB and URLLC and the relation between splitting power fraction and sum-rate of URLLC in different channel conditions are shown and analysed. The rate region of two URLLC users in eMBB and URLLC coexistence is simulated to give a clearer analysis of what is a suitable scenario for RSMA. For eMBB and mMTC coexistence, the achievable pairs of eMBB rate and mMTC arrival rate and the relation between splitting power fraction and mMTC arrival rate are presented and analysed. The suitable situations for RSMA are also discussed in this section.

3.5.1 eMBB and URLLC Coexistence

The simulation results when $n_U = 2$ are shown. Rate regions (r_B^{sum}, r_U^{sum}) are presented in Fig. 3.3 and Fig. 3.4. There is one eMBB user and two URLLC users, and the total number of frequency resources is $F = 10$. Note that to make the figures more intuitive these rate regions are presented as enclosed, and each achievable rate pair is denoted by a marker.

In Fig. 3.3, $\Gamma_B = 10$ dB, $\Gamma_U = 20$ dB, $\epsilon_B = 10^{-3}$, $\epsilon_U = 10^{-5}$. This result shows that NOMA always outperforms OMA, and RSMA outperforms these two schemes. In this scenario, $\Gamma_U > \Gamma_B$, non-orthogonal slicing schemes can leverage the difference between channel gains of URLLC and eMBB to perform SIC, so NOMA and RSMA outperform OMA, especially at a higher rate of eMBB, and RSMA can always achieve a larger rate region than NOMA. The gain of RSMA comes from the decoding order flexibility.

By allocating power to split streams appropriately, both URLLC users can achieve a higher r_U and then a higher sum-rate r_U^{sum} can be obtained. We can imagine a simple situation to get the intuition. Assuming that the channel gains of URLLC user 1 and URLLC user 2 are 21 dB and 19 dB, respectively. The target SNR of eMBB is assumed to 10 dB. If NOMA is applied and the decoding order is URLLC user 1, URLLC user 2 and eMBB, we can easily calculate the rate for URLLC user 1 and URLLC user 2 are 1.26 (bits/s/Hz) and 3.04 (bits/s/Hz). Obviously, r_U should not be higher than 1.26 (bits/s/Hz) if we want both the two users to be decoded. If RSMA is applied and URLLC user 1 splits its message into two streams and as we introduced splitting power fraction to refer to the power fraction allocated to the first stream, we assume the splitting power fraction is 0.8, we can use the decoding order discussed in Section 3.3.3. The reason is this decoding order allows users to achieve all the points of the rate region, which is also explained in Section 2.1, and if the two split streams are decoded successively, it would be as the same as NOMA. Therefore, to maintain the flexibility of RSMA, the two URLLC users are decoded alternatively. We note that different splitting power fraction has different performances, and here we use 0.8 to just give a toy example to show that RSMA can improve the performance. Then, we calculate the rate for URLLC user 1 and URLLC user 2, and the results are 2.62 (bits/s/Hz) and 1.68 (bits/s/Hz), respectively. Now r_U can be 1.68 (bits/s/Hz), so r_U^{sum} also increases.

In Fig. 3.4, $\Gamma_B = 20$ dB, $\Gamma_U = 10$ dB, $\epsilon_B = 10^{-3}$, $\epsilon_U = 10^{-5}$. This simulation result shows that OMA outperforms NOMA except when r_{sum}^B is relatively high. RSMA can outperform NOMA for the same reason in the scenario $\Gamma_B = 10$ dB, $\Gamma_U = 20$ dB which is shown in Fig. 3.3, and outperform OMA when r_B^{sum} is in the range from 0–5 (bits/s/Hz) and in the range from 25–40 (bits/s/Hz). When $\Gamma_B > \Gamma_U$, it is more difficult to satisfy the reliability requirements of URLLC devices when interference from the eMBB device exists, so OMA is preferable in this scenario. OMA can allocate more frequency channels to URLLC, while in NOMA and RSMA, URLLC users tolerate the interference from eMBB user during the SIC procedure. However, if the

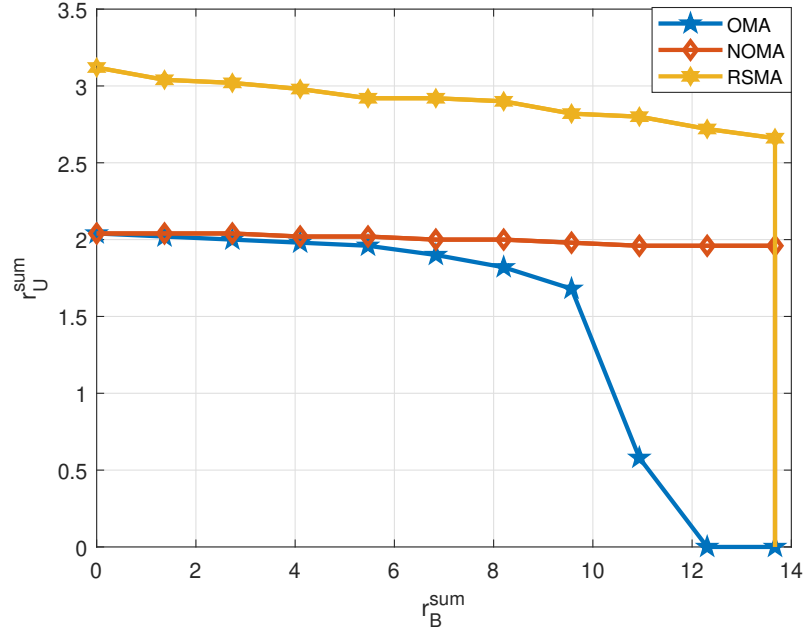


Fig. 3.3. Rate region (r_B^{sum}, r_U^{sum}) , where r_B^{sum} , r_U^{sum} are the sum-rate of eMBB and URLLC, respectively. $\Gamma_B = 10$ dB, $\Gamma_U = 20$ dB, $F = 10$, $\epsilon_B = 10^{-3}$, $\epsilon_U = 10^{-5}$.

aim is to achieve a relatively high r_B^{sum} , NOMA can perform better, because eMBB device can take advantage of frequency diversity. Similarly, RSMA can outperform OMA, and the improvement made by RSMA is more obvious when interference from eMBB device is relatively low, and this is also shown in Fig. 3.3.

Fig. 3.5 and Fig. 3.6 present the relation between splitting power fraction β and the sum-rate of URLLC r_U^{sum} . URLLC users occupy $F_U = 5$ frequency resources in the OMA case, and the eMBB user occupies another $F - F_U = 5$ frequency resources. In both NOMA and RSMA cases, the eMBB rate is fixed at the rate which is the same as the one in OMA case. Since there is no power splitting in OMA and NOMA, the lines corresponding to OMA and NOMA are horizontal lines. In Fig. 3.5, $\Gamma_B = 10$ dB, $\Gamma_U = 20$ dB, and r_U^{sum} achieves the maximal point at $\beta = 0.95$. In this scenario, the URLLC channel condition is better than the eMBB one, so RSMA has more space to adjust the splitting power fraction. However, in Fig. 3.6, $\Gamma_B = 20$ dB, $\Gamma_U = 10$ dB, r_U^{sum} increases as β increases, and it achieves the maximum point at $\beta = 1$, which means RSMA boils down to NOMA. The reason is that due to the relatively high

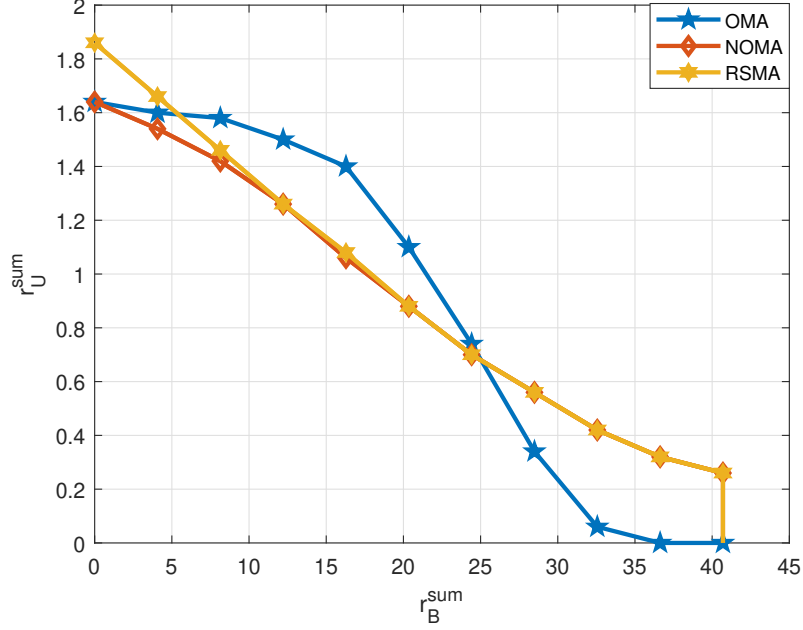


Fig. 3.4. Rate region (r_B^{sum}, r_U^{sum}) , where r_B^{sum} , r_U^{sum} are the sum-rate of eMBB and URLLC, respectively. $\Gamma_B = 20$ dB, $\Gamma_U = 10$ dB, $F = 10$, $\epsilon_B = 10^{-3}$, $\epsilon_U = 10^{-5}$.

eMBB interference, it is difficult for URLLC user 2 to achieve r_U while having the interference from a split stream of URLLC user 1, so splitting power does not make that much difference in this situation. This explains why RSMA and NOMA have the same performance when the rate of eMBB is relatively high in Fig. 3.4.

The reason why RSMA outperforms NOMA is that it can provide a more flexible SIC decoding order to achieve all the boundary points of the rate region theoretically. Thus, a higher r_U and r_U^{sum} could be found in RSMA. In eMBB and URLLC coexistence, the rate regions of two URLLC users in RSMA and NOMA are shown in Fig. 3.7, and the eMBB rate is the same as the rate in Fig. 3.5 and Fig. 3.6. In Fig. 3.7, the boundary points are the average of achievable rate pairs which satisfy the reliability requirements. NOMA without time-sharing and joint encoding/decoding can only achieve the points in the region bounded by the red line, while RSMA can achieve the points in the region bounded by the blue line, which includes the region of NOMA. The essence is that SINR of both URLLC users is adjustable by changing the splitting power fraction, so that RSMA can achieve all the points on the diagonal

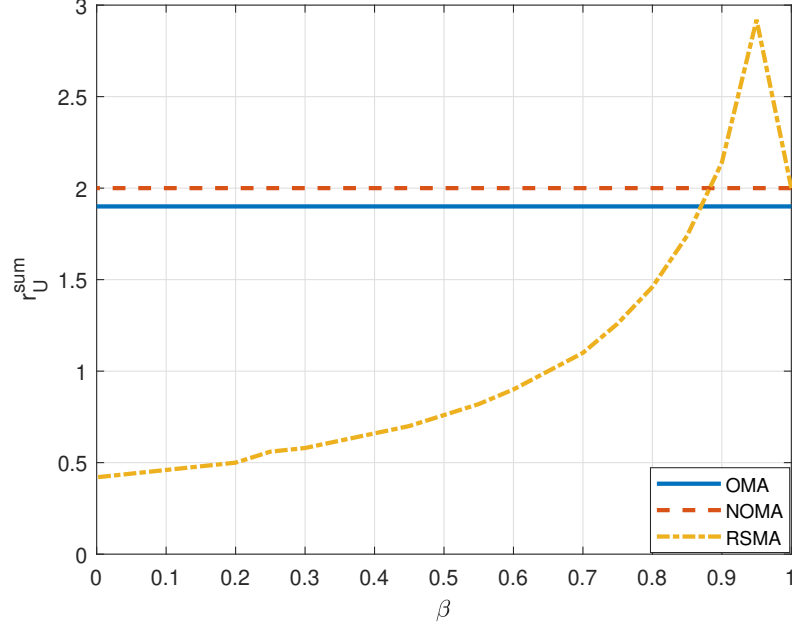


Fig. 3.5. r_U^{sum} versus power factor β . $\Gamma_B = 10$ dB, $\Gamma_U = 20$ dB, $F = 10$, $\epsilon_B = 10^{-3}$, $\epsilon_U = 10^{-5}$.

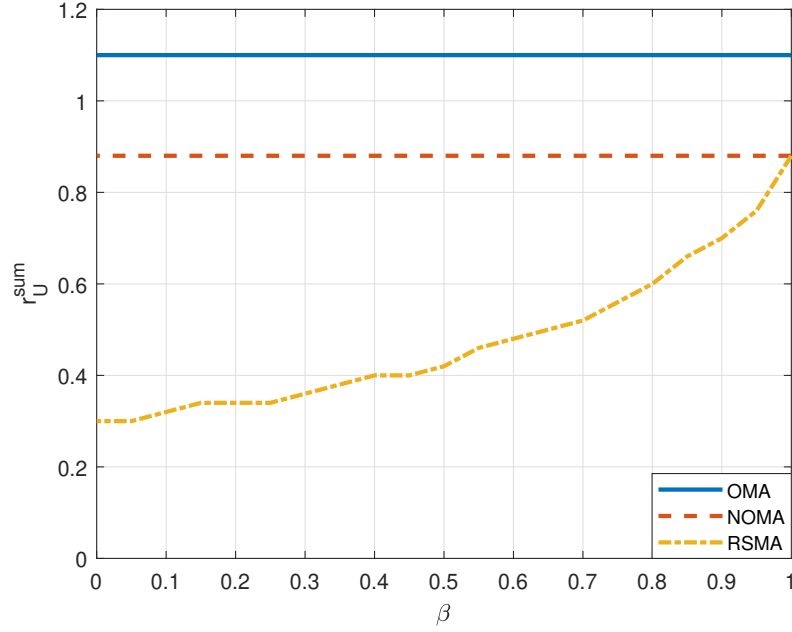


Fig. 3.6. r_U^{sum} versus power factor β . $\Gamma_B = 20$ dB, $\Gamma_U = 10$ dB, $\epsilon_B = 10^{-3}$, $\epsilon_U = 10^{-5}$.

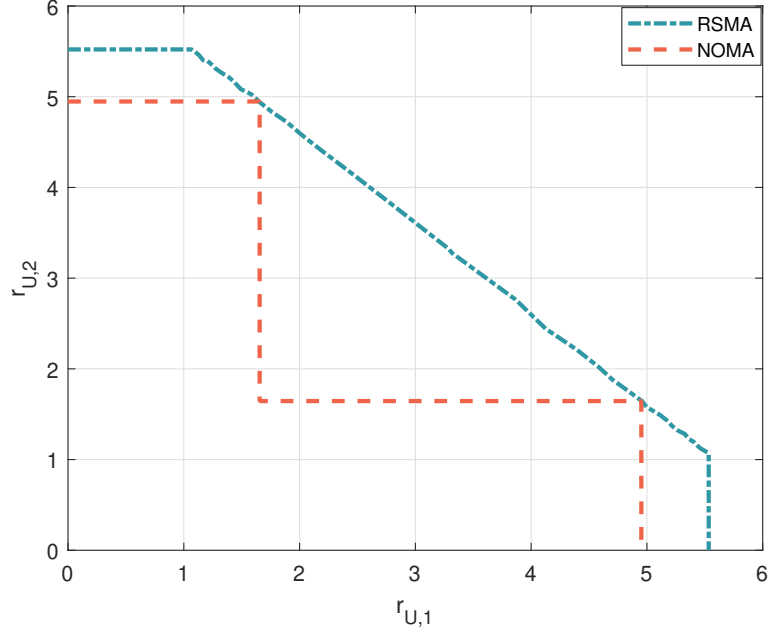
line instead of just two corner points in the rate region, and this can be seen as the user with higher rate compensates the user with lower rate to achieve the desired rate. When $\Gamma_B = 10$ dB and $\Gamma_U = 20$ dB, the SINR of URLLC users is high, the two corner points of NOMA are distant and the difference of regions of NOMA and RSMA is big,

which is shown in Fig. 3.7(a). In this situation, RSMA has more room to adjust the splitting power fraction to obtain the points on the diagonal line, and a higher r_U can be found. However, when $\Gamma_B = 20$ dB and $\Gamma_U = 10$ dB and the SINR of URLLC users is low, the two corner points of NOMA can be very close due to the high interference from eMBB, so the difference between NOMA and RSMA regions becomes smaller, which is shown in Fig. 3.7(b). In this scenario, RSMA only has a small room to play its role, and the user with a higher rate also struggles to satisfy the rate requirement, so it is more difficult to compensate the user with a lower rate. Thus, RSMA obtains better performance than both OMA and NOMA when r_B^{sum} is relatively low, while obtains the same performance as NOMA when r_B^{sum} is high.

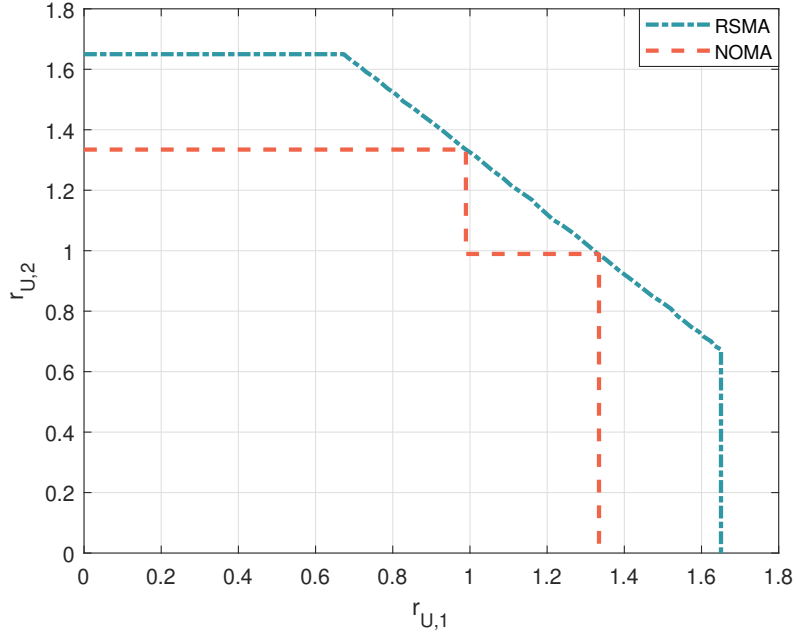
3.5.2 eMBB and mMTC Coexistence

The simulation results of achievable pairs (r_B, λ_M) are presented. As we assume that the number of active mMTC devices follows Poisson distribution, $n_M \sim \text{Poisson}(\lambda_M)$, where λ_M is the mean number, we refer λ_M to as mMTC arrival rate. λ_M is derived by exhaustive search, which means that λ_M keep increasing in each simulation run until the greatest λ_M satisfying (3.30) is found. In Fig. 3.8, $\Gamma_B = 20$ dB, $\Gamma_M = 5$ dB, $r_M = 0.04$ (bits/s/Hz), $\epsilon_B = 10^{-3}$, $\epsilon_M = 10^{-1}$. This result shows that when $r_B > 1.8$ (bits/s/Hz), OMA can achieve higher λ_M than NOMA; while when $0.5 < r_B < 1.8$ (bits/s/Hz), NOMA achieves higher λ_M . RSMA always outperforms NOMA and outperforms OMA when $0.5 < r_B < 2.5$ (bits/s/Hz), but it still cannot achieve the region that OMA achieves when $r_B > 2.5$ (bit/s/Hz). The curve of OMA is approximately linear, while for NOMA and RSMA the curves drop dramatically at $r_B = 1.5$ (bit/s/Hz) and $r_B = 2$ (bit/s/Hz), respectively.

When r_B is relatively low, i.e. $0 - 0.5$ (bits/s/Hz), the performances of these three schemes are very similar, because for NOMA and RSMA the interference from eMBB device is relatively low, so it almost does not affect mMTC traffic, and for OMA, a



(a)



(b)

Fig. 3.7. The rate regions $(r_{U,1}, r_{U,2})$ of two URLLC users in eMBB and URLLC coexistence scenario. (a) is the rate region when $\Gamma_B = 10$ dB, $\Gamma_U = 20$ dB, $\epsilon_B = 10^{-3}$ and $\epsilon_U = 10^{-5}$. (b) is the rate region when $\Gamma_B = 20$ dB, $\Gamma_U = 10$ dB, $\epsilon_B = 10^{-3}$ and $\epsilon_U = 10^{-5}$.

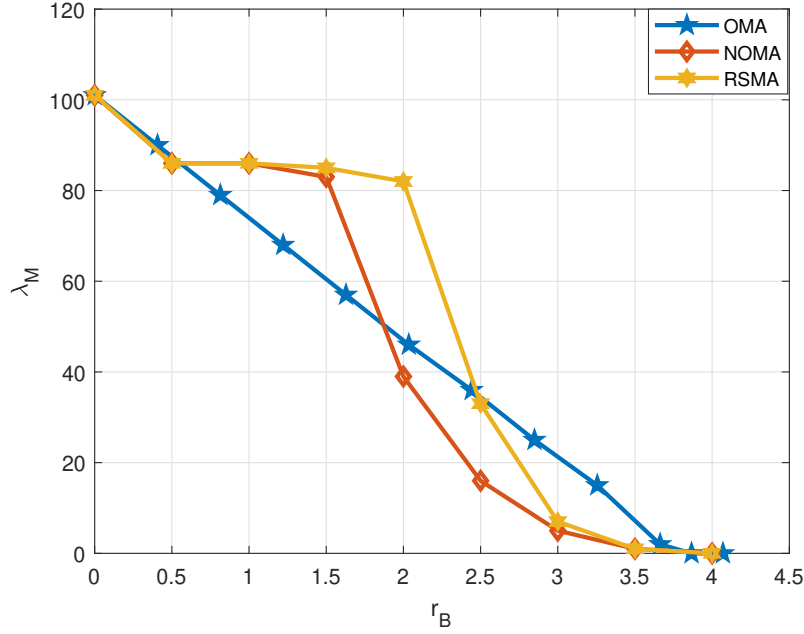


Fig. 3.8. Achievable pairs (r_B, λ_M) , where r_B , λ_M are the rate of eMBB and the arrival rate of mMTC traffic, respectively. $\Gamma_B = 20$ dB, $\Gamma_M = 5$ dB, $r_M = 0.04$ (bits/s/Hz), $\epsilon_B = 10^{-3}$, $\epsilon_M = 10^{-1}$.

large time fraction is allocated to mMTC traffic and eMBB only occupies a small one. When r_B is medium, i.e. $0.5 - 2$ (bits/s/Hz), RSMA and NOMA outperform OMA. At this eMBB rate, BS can leverage the differences between channel gains and reliability requirements of eMBB and mMTC devices when NOMA or RSMA is applied, but OMA cannot leverage these differences. When eMBB rate is relatively high, i.e. $2 - 4$ (bits/s/Hz), OMA outperforms NOMA, and it also outperforms RSMA when $r_B > 2.5$ (bits/s/Hz). eMBB device will cause high interference to mMTC devices due to its relatively high rate, and almost no mMTC devices can be decoded before the eMBB device. Because of the power constraint of the eMBB device, the eMBB device cannot tolerate much interference from mMTC devices, so λ_M drops to satisfy ϵ_M . While for OMA, eMBB and mMTC traffic are isolated, so mMTC will not be affected by eMBB so much compared to NOMA and RSMA.

RSMA mainly outperforms NOMA at a medium eMBB rate, i.e. $1.5 - 3$ (bits/s/Hz). At this rate region, some mMTC devices must be decoded after eMBB due to the insufficient SINR, so if eMBB cannot be decoded, the decoding procedure is 'lodged'

and then terminates. As analysed in the previous section, using RSMA can let part of the eMBB interference be cancelled, and in this way, the decoding procedure avoids 'lodged' to some extent, so RSMA has gain in this situation. For low eMBB rate and high eMBB rate, RSMA does not make much difference, because at a low rate almost all mMTC devices can be decoded before the eMBB device, and at a high rate mMTC devices must be decoded after eMBB device, and RSMA boils down to NOMA.

We note that the simulation result in Fig. 3.8 is in system level and perfect SIC is considered to study the fundamental performances. Recalling the error probability of mMTC formulated in (2.28), it measures the fraction of incorrectly detected active users, so λ_M is not the number of decoded mMTC users, but the greatest number to ensure that at least $\lambda_M(1 - \epsilon_M)$ mMTC users can be decoded. As such, the channel disparity could be huge between users, and some mMTC users may have very high channel gains. Since mMTC does not expect a high data rate like eMBB, the mMTC users with high channel gains could tolerate interference from many other users, so that in Fig. 3.8 mMTC can have a high arrival rate. However, in the physical world, SIC cannot be performed perfectly and the performance of SIC is also limited by the modulation index because a higher modulation index requires a more careful power allocation strategy. Therefore, in the real world, it is more likely to support a lower arrival rate λ_M .

Fig. 3.9 shows the relation between mMTC arrival rate λ_M and power factor β when eMBB rate is fixed at 2 (bits/s/Hz). Since OMA and NOMA do not split power, the corresponding curves are horizontal. For RSMA, the arrival rate achieves the maximum point at $\beta = 0.45$, and is the same as NOMA when $\beta = 0$ and $\beta = 1$. As shown in Fig. 3.8, $r_B = 2$ (bits/s/Hz) is a relatively high rate, and either mMTC devices or eMBB device may not have enough SINR to be decoded, which causes decoding procedure to terminate. By adjusting the splitting power fraction of the eMBB device, the SINR of mMTC devices is also adjusted, so that more mMTC devices can

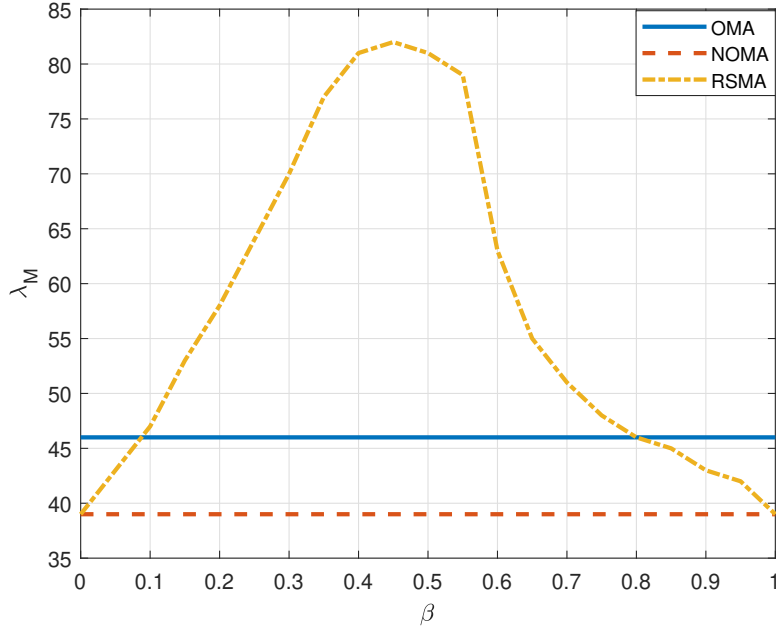


Fig. 3.9. Arrival rate λ_M versus power factor β . $\Gamma_B = 20$ dB, $\Gamma_M = 5$ dB, $r_M = 0.04$ (bits/s/Hz), $\epsilon_B = 10^{-3}$, $\epsilon_M = 10^{-1}$.

be decoded. Thus, λ_M increases as β increases till $\beta = 0.45$, and then decreases to the point which is as same as NOMA at $\beta = 1$.

3.6 Summary

In this chapter, we investigate eMBB and URLLC coexistence and eMBB and mMTC coexistence with OMA, NOMA and RSMA. For RSMA scheme designs, it is important to decide properly which user splits the message and design the power allocations. Given the different requirements of eMBB, URLLC and mMTC, different splitting strategies are used. For eMBB and URLLC coexistence, RSMA is applied between URLLC users to fulfil the requirements of eMBB and URLLC. The performances in different channel conditions are simulated and analysed, and then we discuss the suitable scenario for RSMA. For eMBB and mMTC coexistence, a novel RSMA scheme is proposed. RSMA is applied between eMBB and mMTC because of service requirements and the nature of mMTC arrival process, and the performance is simulated

and analysed.

According to the simulation results, RSMA always outperforms NOMA in terms of achievable pairs for both coexistence scenarios. RSMA and NOMA can achieve the achievable pairs that OMA cannot; while OMA can also achieve the pairs that RSMA or NOMA cannot. The advantages of RSMA are that it can leverage the differences between channel gains and reliability requirements, and it achieves a larger rate region. The splitting power fraction of RSMA is a very significant factor to achieve the optimal sum-rate, and this chapter gives insights into how the splitting power fraction can impact the performances. For eMBB and URLLC coexistence, the lower the interference from eMBB, the higher the improvement RSMA can make, because RSMA has more room to adjust the power allocation and obtain the optimal rate pairs while guaranteeing the service requirements. For eMBB and mMTC coexistence, due to splitting the messages, it is possible for RSMA to let part of the eMBB interference be cancelled, so that it can support a higher arrival rate of mMTC for a given reliability requirement when the splitting power fraction is set properly. This chapter concludes that RSMA is a promising technique for network slicing to fulfil the requirements of different users efficiently.

Chapter 4

RSMA for Coexistence of Semantic and Bit Communications

In 6G, the demands for capacity and connectivity will increase dramatically to meet the requirements of emerging services for both humans and machines. Semantic communication has shown great potential because of its efficiency, and suitability for users who only care about the semantic meaning. But bit communication is still needed for users requiring original messages. Therefore, there will be a coexistence of semantic and bit communications in future networks. This motivates us to explore how to allocate resources in such a coexistence scenario. We investigate different uplink MA schemes for the coexistence of semantic users and a bit user, namely OMA, NOMA and RSMA. We characterize the rate regions achieved by those MA schemes. The simulation results show that RSMA always outperforms NOMA and has better performance in high semantic rate regimes compared to OMA. We find that RSMA scheme design, rate region, and power allocation are quite different in the coexistence scenario compared to the bit-only communication, primarily due to the need to consider the understandability in semantic communications. Interestingly, in contrast to bit-only communications where RSMA is capacity achieving without any need for time

sharing, in the coexistence scenario, time sharing helps enlarge RSMA rate region.

4.1 Introduction

Semantic communication has attracted attention because of its efficiency and suitability for users who only care the semantic meaning. Future networks are expected to support both semantic communications and bit communications, so how to allocate resources between two types of communications is critical and it drives us to study MA schemes for a semantic-user-and-bit-user coexistence scenario. In light of the literature we reviewed in Section 2.3, an interesting question arises: since RSMA has already shown lots of good properties in bit-communication and in downlink semantic communication, will RSMA still hold its advantages in this coexistence scenario? This chapter tackles this question by investigating MA schemes for a semantic-user-and-bit-user coexistence scenario in the uplink, simply referred to as coexistence in this chapter. Unlike bit communications, semantic communications also require understandability, which means that the meaning of received messages should be unchanged. Besides, the definition of transmission efficiency is also different from bit communications. Therefore, these specificities of semantic communications should be considered in the MA scheme designs for a coexistence scenario, and the RSMA design for the coexistence scenario cannot be simply obtained by extending the existing work of bit communications. Although there are works about downlink RSMA for semantic communications, their results cannot be used directly to a uplink coexistence scenario because of the differences between uplink RSMA and downlink RSMA [17, 36]. These raise the challenges of uplink RSMA design for the coexistence scenario, i.e. the decoding order of split streams and power allocation. This chapter focuses on the coexistence of multiple semantic users and one bit user, and we consider semantic communications of text transmission, whose framework is based

on [74]. The contributions of this chapter are below.

- This work investigates different MA schemes for the coexistence scenario in the uplink, namely OMA, NOMA and RSMA. To our best knowledge, it is the first time to study the MA schemes for the coexistence scenario in the uplink. In the coexistence scenario, besides the requirement of bit users, the transmission efficiency and understandability of semantic users also need to be considered. Therefore, it is necessary to take into account those new factors when designing the MA schemes for the coexistence scenario. Besides commonly used OMA and NOMA, we propose a RSMA scheme for the coexistence scenario.
- We characterize the rate region of these three schemes. While in conventional bit communication, the rate region is the set of bit rate pairs that the transmitters can achieve simultaneously, in this coexistence scenario the rate region is the set of achievable semantic rate and bit rate pairs. To obtain the boundary of the rate regions, we fix the semantic rate and then find the maximum achievable bit rate. For OMA, we consider FDMA and transform this problem into finding the minimum bandwidth required by semantic users. For NOMA and RSMA, we propose a SCA algorithm to solve the problems.
- We simulate the performances of OMA, NOMA and RSMA, and discuss the suitable scenarios for RSMA according to observations and analysis. We show the rate regions of these three schemes with different numbers of semantic users and different distances between semantic users and BS. These results show that in the coexistence scenario, RSMA has a larger rate region than NOMA because of its flexibility in the decoding procedure. Although RSMA cannot outperform OMA in low semantic rate regimes because an interference-free resource allocation strategy is preferable, in most situations RSMA has better performances, especially in high semantic rate regimes. Then we present the relation between the number of semantic users and the achievable bit rate. The

superiority of RSMA is more obvious as the number of semantic users increases. We also simulate the rate regions of different sentence similarity thresholds, and we show that the improvement of RSMA is more salient when the sentence similarity threshold is high.

- We summarize the main differences between RSMA design for bit communications and for the coexistence scenario, and this summary of how RSMA design changes may bring insights for the future related studies. Firstly, in the RSMA scheme design, our RSMA for coexistence does not split the semantic users and allows the bit user to split multiple times. This contrasts with conventional RSMA for bit communications where all users, except one, are split into two parts. We use this scheme because splitting the semantic user may hinder the understandability and splitting bit users multiple times can provide more flexibility in the decoding procedure. Secondly, we observe that the rate region of RSMA for the coexistence is quite different from the one of bit communications due to the constraints on the understandability and transmission efficiency of semantic users. Then we contrast the power allocation strategies of RSMA for the coexistence and RSMA for bit communications to assess the impact of semantic communications on the RSMA designs further. Finally, we find that in the coexistence RSMA cannot always outperform OMA due to the need to account for the understandability of semantic users, while in bit communications RSMA always performs better than OMA and NOMA. However, we can use time sharing between OMA and RSMA to obtain the largest rate region.

4.2 System Model

We consider that there are multiple semantic users and a bit user and they communicate to a common BS. Bit users are interested in the original messages, while semantic

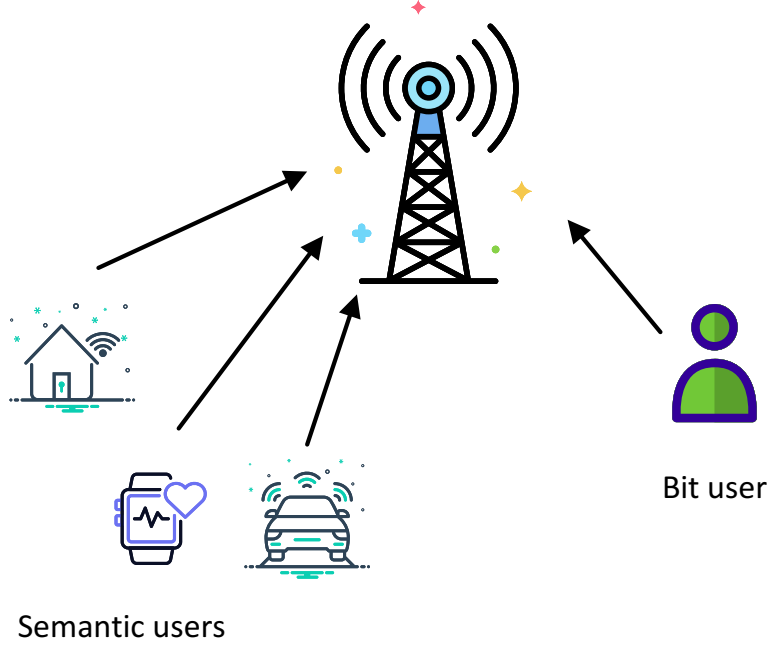


Fig. 4.1. An illustration of semantic users and bit user coexistence.

users only care about the meaning and most of them are machine-type users. Therefore, the number of semantic users can be much more than the number of bit users, so the scenario that a bit user and multiple semantic users can grasp the essence. An illustration is shown in Fig. 4.1. The BS and the users are single-antenna. We use N_s to denote the number of semantic users, $U_{s,j}$ denote the semantic users, where $j \in \mathcal{J}, \mathcal{J} = \{1, 2, \dots, N_s\}$, and U_b denote the bit user. Let $h_{s,j} \in \mathbb{C}^{1 \times 1}$ and $h_b \in \mathbb{C}^{1 \times 1}$ denote the channel coefficients for $U_{s,j}$ and U_b , respectively. The transmission power of $U_{s,j}$ and U_b are $p_{s,j}$ and p_b , respectively, and the maximum transmission power is P for all users. The total available bandwidth is w . Let $x_{s,j}$ denote the normalized semantic symbol of $U_{s,j}$ and x_b denote the normalized conventional symbol of U_b . Since multiple users share limited resources, designing an appropriate MA scheme would be important to this coexistence scenario. We first study two commonly used MA techniques, OMA and NOMA. For OMA, here we will consider FDMA. Then we propose a resource allocation scheme relying on RSMA for this coexistence scenario. To evaluate the perfor-

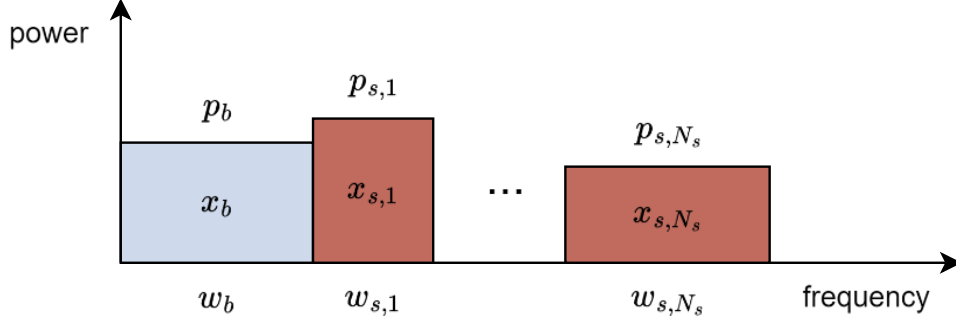


Fig. 4.2. The illustration of FDMA.

mances of different MA schemes, we investigate the rate regions of these three schemes. Different from the rate region in conventional bit communication, here the rate region is a set of semantic rate and bit rate pairs. We assume that CSI is known at the BS to study the fundamental performance of the three schemes, and the BS will inform the resource allocation strategies to all the users during the connection setup process. The semantic users require the same service so they should satisfy the same requirements.

4.2.1 FDMA

In FDMA, the bit user and the semantic users are allocated isolated frequency bands. Let $w_{s,j}$ and w_b denote the bandwidths allocated to $U_{s,j}$ and U_b , respectively. The illustration of FDMA is shown in Fig. 4.2. The received signal can be represented as

$$y = \sum_{j=1}^{N_s} h_{s,j} x_{s,j} + h_b x_b + n, \quad (4.1)$$

where n is the additive white noise. The SINR of the bit user is

$$\rho_b = \frac{p_b |h_b|^2}{w_b N_0}, \quad (4.2)$$

where N_0 is the noise spectral power density. The achievable bit rate is

$$r_b^F = w_b \log_2(1 + \rho_b). \quad (4.3)$$

The SINR of $U_{s,j}$ is

$$\rho_{s,j} = \frac{p_{s,j} |h_{s,j}|^2}{w_{s,j} N_0}. \quad (4.4)$$

According to the approximation of sentence similarity in (2.32) which was introduced by [75], we have

$$\varepsilon(\rho_{s,j}) = A_{k,1} + \frac{A_{k,2} - A_{k,1}}{1 + e^{-(C_{k,1}\rho_{s,j} + C_{k,2})}}. \quad (4.5)$$

The sentence similarity should be higher than a threshold S^{th} so that the semantic message can be decoded [74]. S^{th} satisfies $A_{k,1} \leq S^{th} \leq A_{k,2}$ because $A_{k,2}$ and $A_{k,1}$ are the upper bound and the lower bound of the sentence similarity, respectively. Thus,

$$\begin{aligned} \varepsilon(\rho_{s,j}) &\geq S^{th} \\ \rho_{s,j} &\geq -\frac{1}{C_{k,1}} \left[C_{k,2} + \ln \left(\frac{A_{k,2} - A_{k,1}}{S^{th} - A_{k,1}} - 1 \right) \right] = \gamma_{sem}, \end{aligned} \quad (4.6)$$

which means that the SINR of the semantic users should be higher than γ_{sem} . Then, the achievable semantic rate for $U_{s,j}$ is

$$S_j^F = \frac{w_j I}{K L} \varepsilon(\rho_{s,j}). \quad (4.7)$$

Therefore, the rate region of coexistence with FDMA is

$$\begin{aligned} C^F \triangleq & \bigcup \left\{ (S^F, R^F) : S^F \leq S_j^F, R_b^F \leq r_b^F, \varepsilon(\rho_{s,j}) \geq S^{th}, \right. \\ & \left. p_b \leq P, p_{s,j} \leq P, w_b + \sum_{j=1}^{N_s} w_{s,j} \leq w, j \in \mathcal{J} \right\}. \end{aligned} \quad (4.8)$$

To characterize the rate region, we can fix an achievable semantic rate S , and find the highest achievable rate r_b^F . In this way, each semantic rate has a corresponding maximum r_b^F , so the rate region can be obtained. Therefore, for a given S^{th} and semantic rate S , a problem that maximizes the bit rate can be formulated as below:

$$\max_{\mathbf{w}_s, w_b, \mathbf{p}_s, p_b} w_b \log_2 \left(1 + \frac{p_b |h_b|^2}{w_b \sigma_n^2} \right), \quad (4.9a)$$

$$\text{s.t. } \frac{w_{s,j} I}{KL} \varepsilon \left(\frac{p_{s,j} |h_{s,j}|^2}{w_{s,j} N_0} \right) \geq S, \forall j \in \mathcal{J}, \quad (4.9b)$$

$$\frac{p_{s,j} |h_{s,j}|^2}{w_{s,j} N_0} \geq \gamma_{sem}, \forall j \in \mathcal{J}, \quad (4.9c)$$

$$p_b \leq P, \quad (4.9d)$$

$$p_{s,j} \leq P, \forall j \in \mathcal{J}, \quad (4.9e)$$

$$\sum_{j \in \mathcal{J}} w_{s,j} + w_b \leq w, \quad (4.9f)$$

where $\mathbf{w}_s = [w_{s,1}, w_{s,2}, \dots, w_{s,N_s}]$, $\mathbf{p}_s = [p_{s,1}, p_{s,2}, \dots, p_{s,N_s}]$, and P is the power constraint. In the optimal solution, $w_b = w - \sum_{j \in \mathcal{J}} w_{s,j}$ and the bit user will use full power to transmit, which means $p_b = P$. The problem can be transformed to

$$\min_{\mathbf{w}_s, \mathbf{p}_s} \sum_{j \in \mathcal{J}} w_{s,j}, \quad (4.10a)$$

$$\text{s.t. } (4.9b), (4.9c) \text{ and } (4.9e), \quad (4.10b)$$

$$\sum_{j \in \mathcal{J}} w_{s,j} \leq w. \quad (4.10c)$$

To solve this problem, we first consider that with a given $p_{s,j}$, what is the smallest $w_{s,j}$ satisfying (4.9b) and (4.9c). Then we consider that which $p_{s,j}$ can give the smallest $w_{s,j}$. We use the following lemmas to obtain the solution.

Lemma 1: The semantic rate of $U_{s,j}$ increases when $w_{s,j}$ increases. For a given S and a $p_{s,j}$, the smallest $w_{s,j}$ that satisfies (4.9b) is $w_{s,j}^*$ and $\frac{w_{s,j}^* I}{KL} \varepsilon \left(\frac{p_{s,j} |h_{s,j}|^2}{w_{s,j}^* N_0} \right) = S$.

Proof. We use $f(w_{s,j})$ to denote the achievable semantic rate and we note that ρ in (2.32) is a decibel number, so

$$\begin{aligned} f(w_{s,j}) &= \frac{w_{s,j}I}{KL} \left\{ A_{k,1} + \frac{A_{k,2} - A_{k,1}}{1 + e^{-\left[10C_{k,1}\lg\left(\frac{p_{s,j}|h_{s,j}|^2}{w_{s,j}N_0}\right) + C_{k,2}\right]}} \right\}. \end{aligned} \quad (4.11)$$

The partial derivative of f with respect to the $w_{s,j}$ is

$$\begin{aligned} \frac{\partial f(w_{s,j})}{\partial w_{s,j}} &= \frac{I}{KL} \left[A_{k,1} + \frac{A_{k,2} - A_{k,1}}{1 + \left(\frac{p_{s,j}|h_{s,j}|^2}{w_{s,j}N_0}\right)^{-\frac{10C_{k,1}}{\ln 10}} e^{-C_{k,2}}} \right] \\ &\quad - \frac{I}{KL} \left\{ (A_{k,2} - A_{k,1}) \frac{e^{-C_{k,2}} \frac{10C_{k,1}}{\ln 10} \left(\frac{p_{s,j}|h_{s,j}|^2}{w_{s,j}N_0}\right)^{-\frac{10C_{k,1}}{\ln 10}}}{\left[1 + \left(\frac{p_{s,j}|h_{s,j}|^2}{w_{s,j}N_0}\right)^{-\frac{10C_{k,1}}{\ln 10}} e^{-C_{k,2}}\right]^2} \right\} \\ &= \frac{I}{KL} A_{k,1} \\ &\quad + \frac{I(A_{k,2} - A_{k,1})}{KL} \frac{1 + \left(\frac{p_{s,j}|h_{s,j}|^2}{w_{s,j}N_0}\right)^{-\frac{10C_{k,1}}{\ln 10}} e^{-C_{k,2}} \left(1 - \frac{10C_{k,1}}{\ln 10}\right)}{\left[1 + \left(\frac{p_{s,j}|h_{s,j}|^2}{w_{s,j}N_0}\right)^{-\frac{10C_{k,1}}{\ln 10}} e^{-C_{k,2}}\right]^2}. \end{aligned} \quad (4.12)$$

The first term and the denominator of the second term are positive in (4.12), and we consider the numerator of the second term. $\frac{p_{s,j}|h_{s,j}|^2}{w_{s,j}N_0}$ should be no smaller than γ_{sem} , otherwise, it will violate (4.9c). Since I/L , $A_{k,1}$, $A_{k,2}$, $C_{k,1}$ and $C_{k,2}$ are constant, and specifically $C_{k,1}$ is positive, we can obtain the smallest value of (4.12) when $\frac{p_{s,j}|h_{s,j}|^2}{w_{s,j}N_0} = \gamma_{sem}$, which is a positive number. Thus, the semantic rate increases when $w_{s,j}$ increases, so the smallest $w_{s,j}$ is $w_{s,j}^*$ where $\frac{w_{s,j}^*}{K} \varepsilon \left(\frac{p_{s,j}|h_{s,j}|^2}{w_{s,j}^*N_0}\right) = S$. The proof of Lemma 1 is completed. $\square$

From (4.9b) and Lemma 1, we can obtain that

$$w_{s,j} \geq w_{s,j}^* = w_{s,j}^l, \quad (4.13)$$

and $w_{s,j}^*$ satisfies that

$$\frac{w_{s,j}^* I}{KL} \varepsilon \left(\frac{p_{s,j} |h_{s,j}|^2}{w_{s,j}^* N_0} \right) = S. \quad (4.14)$$

From (4.9c), we can obtain that

$$w_{s,j} \leq \frac{p_{s,j} |h_{s,j}|^2}{\gamma_{sem} N_0} = w_{s,j}^h. \quad (4.15)$$

If (4.10a) is feasible, a feasible solution $w_{s,j}$ should satisfy (4.9b) and (4.9c), which means $w_{s,j} \in [w_{s,j}^l, w_{s,j}^h]$ and $w_{s,j}^h \geq w_{s,j}^l$. In fact, $w_{s,j}^h$ is always larger than $w_{s,j}^l$. (4.15) can be transformed to

$$p_{s,j} |h_{s,j}|^2 = w_{s,j}^h \gamma_{sem} N_0, \quad (4.16)$$

and we substitute (4.16) into (4.14), we have

$$\frac{w_{s,j}^* I}{KL} \varepsilon \left(\frac{w_{s,j}^h \gamma_{sem}}{w_{s,j}^*} \right) = S. \quad (4.17)$$

Therefore, $w_{s,j}^h \geq w_{s,j}^* = w_{s,j}^l$, otherwise, $\varepsilon \left(\frac{w_{s,j}^h \gamma_{sem}}{w_{s,j}^*} \right) < S^{th}$ will violate (4.9c). Since $w_{s,j}^h$ is always larger than $w_{s,j}^l$, the optimal $w_{s,j}$ is $w_{s,j}^l$. Then the next step is finding the optimal $p_{s,j}$ which gives the least $w_{s,j}^l$. To find the optimal $p_{s,j}$, we use the following lemma.

Lemma 2: The smallest $w_{s,j}^l$ is obtained when $p_{s,j} = P$.

Proof. When $p_{s,j} = P$, $\frac{w_{s,j}^l I}{KL} \varepsilon \left(\frac{P |h_{s,j}|^2}{w_{s,j}^* N_0} \right) = S$. We assume that there exists $w_{s,j}^{**} < w_{s,j}^l$ and $p_{s,j}^{**} < P$, and they satisfy $\frac{w_{s,j}^{**} I}{KL} \varepsilon \left(\frac{p_{s,j}^{**} |h_{s,j}|^2}{w_{s,j}^{**} N_0} \right) = S$ and $\frac{p_{s,j}^{**} |h_{s,j}|^2}{w_{s,j}^{**} N_0} \geq \gamma_{sem}$. Obviously, the semantic rate increases when $p_{s,j}$ increases, so $\frac{w_{s,j}^{**} I}{KL} \varepsilon \left(\frac{p_{s,j}^{**} |h_{s,j}|^2}{w_{s,j}^{**} N_0} \right) < \frac{w_{s,j}^{**} I}{KL} \varepsilon \left(\frac{P |h_{s,j}|^2}{w_{s,j}^{**} N_0} \right)$. From Lemma 1, we have $\frac{w_{s,j}^{**} I}{KL} \varepsilon \left(\frac{P |h_{s,j}|^2}{w_{s,j}^{**} N_0} \right) < \frac{w_{s,j}^l I}{KL} \varepsilon \left(\frac{P |h_{s,j}|^2}{w_{s,j}^* N_0} \right) = S$, which contradicts to

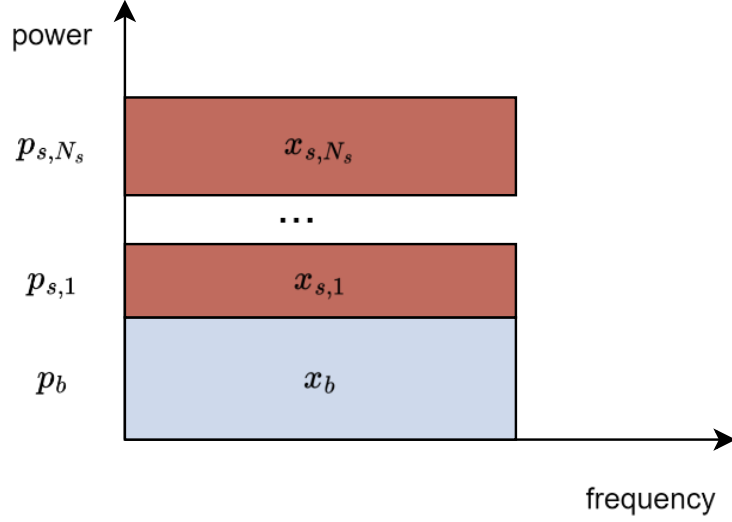


Fig. 4.3. The illustration of NOMA.

the assumptions. The proof of Lemma 2 is completed. $\square$

Therefore, we have the solution of (4.10a), $w_{s,j} = w_{s,j}^*$, where $w_{s,j}^*$ satisfies (4.14) and $p_{s,j} = P$. If $\sum_{j \in \mathcal{J}} w_{s,j} \leq w$, this solution is feasible, otherwise, problem (4.10a) is unsolvable. Thus, we find the highest achievable bit rate for every given S and then we can obtain the rate region.

4.2.2 NOMA

In NOMA, semantic users and the bit user share the same bandwidth as shown in Fig. 4.3. The received signal can be represented as

$$y = \sum_{j \in \mathcal{J}} \sqrt{p_{s,j}} h_{s,j} x_{s,j} + \sqrt{p_b} h_b x_b + n, \quad (4.18)$$

Since semantic users and the bit user will use the same band w , we use σ_n^2 to denote the noise power over w for convenience. BS uses SIC to decode messages, and it is

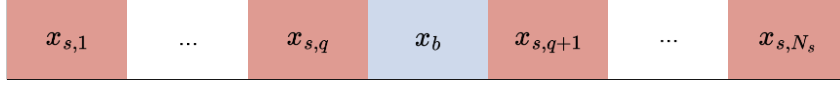


Fig. 4.4. The decoding order of NOMA.

important to design an appropriate decoding order. First, we consider the decoding order among the semantic users. We assume that $U_{s,j}$ is the user with the j -th largest channel gain. From (4.6), we know that for a given sentence similarity threshold S^{th} , the SINR of the semantic users should achieve γ^{sem} , so we decode the users with better channel conditions first, and the decoding order of the semantic users is $U_{s,1}, U_{s,2}, \dots, U_{s,N_s}$. Therefore, when there are N_s semantic users and a bit user, there are $N_s + 1$ decoding orders in total.

We assume that the bit user will be decoded after q semantic users, where $0 \leq q \leq N_s$, and $q = 0$ means that the bit user will be decoded before all the semantic users. An illustration of the decoding order of NOMA is shown in Fig. 4.4. The SINR of the bit user is

$$\rho_b = \frac{p_b |h_b|^2}{\sigma_n^2 + \sum_{i=q+1}^{N_s} p_{s,i} |h_{s,i}|^2}, \quad (4.19)$$

and the achievable rate is

$$r_b^N = w \log_2 \left(1 + \frac{p_b |h_b|^2}{\sigma_n^2 + \sum_{i=q+1}^{N_s} p_{s,i} |h_{s,i}|^2} \right). \quad (4.20)$$

The SINR of $U_{s,j}$ is

$$\rho_{s,j} = \begin{cases} \frac{p_{s,j} |h_{s,j}|^2}{\sigma_n^2 + \sum_{i=j+1}^{N_s} p_{s,i} |h_{s,i}|^2 + p_b |h_b|^2}, & j \leq q, \\ \frac{p_{s,j} |h_{s,j}|^2}{\sigma_n^2 + \sum_{i=j+1}^{N_s} p_{s,i} |h_{s,i}|^2}, & j > q, \end{cases} \quad (4.21)$$

and the achievable semantic rate is

$$S_j^N = \frac{wI}{KL} \varepsilon(\rho_{s,j}). \quad (4.22)$$

The rate region of NOMA is

$$C^N \triangleq \bigcup \left\{ (S^N, R_b^N) : S^N \leq S_j^N, R_b^N \leq r_b^N, \varepsilon(\rho_{s,j}) \geq S^{th}, \right. \\ \left. j \in \mathcal{J}, p_{s,j} \leq P, p_{s,j} \leq P, q = \{0, 1, 2, \dots, N_s\} \right\}. \quad (4.23)$$

Similarly, we fix an achievable S , and find the maximum achievable r_b^N . A problem maximizing the bit rate is formulated as

$$\max_{\mathbf{p}_s, p_b, q} \log_2 \left(1 + \frac{p_b |h_b|^2}{\sigma_n^2 + \sum_{i=q+1}^{N_s} p_{s,i} |h_{s,i}|^2} \right) \quad (4.24a)$$

$$\text{s.t. } \rho_{s,j} \geq \gamma_{sem}, \forall j \in \mathcal{J}, \quad (4.24b)$$

$$\frac{wI}{KL} \varepsilon(\rho_{s,j}) \geq S, \forall j \in \mathcal{J}, \quad (4.24c)$$

$$p_{s,j} \leq P, \forall j \in \mathcal{J}, \quad (4.24d)$$

$$p_b \leq P, \quad (4.24e)$$

$$q = 0, 1, \dots, N_s. \quad (4.24f)$$

From (2.31) and (2.32), we can rewrite (4.24c) to

$$\rho_{s,j} \geq \begin{cases} \gamma_{sem}, & S \leq \frac{S^{th} I}{KL}, \\ -\frac{1}{C_{k,1}} \left[\ln \left(\frac{A_{k,2} - A_{k,1}}{\frac{KL S}{I} - A_{k,1}} - 1 \right) + C_{k,2} \right], & \text{otherwise.} \end{cases} \quad (4.25)$$

(4.25) shows that the semantic rate constraint is equal to a constraint for $\rho_{s,j}$, which is $\rho_{s,j}$ should be higher than a constant and we use γ to denote this constant. Note that the maximum S is $\frac{A_{k,2} I}{KL}$ because $A_{k,2}$ is the upper bound of $\varepsilon(\rho_{s,j})$.

We propose a SCA method to solve the problem (4.24a). Since (4.24a) is non-convex due to the non-convex rate expressions, we introduce the following variables, t , ρ_b and σ_b . Then the problem is transformed to

$$\max_{\mathbf{p}_s, p_b, \rho_b, \sigma_b, q} t \quad (4.26a)$$

$$\text{s.t. } w \log_2(1 + \rho_b) \geq t, \quad (4.26b)$$

$$\frac{p_b |h_b|^2}{\sigma_b} \geq \rho_b, \quad (4.26c)$$

$$\sigma_b^2 + \sum_{i=q+1}^{N_s} p_{s,i} |h_{s,i}|^2 \leq \sigma_b, \quad (4.26d)$$

$$\rho_{s,j} \geq \gamma, \forall j \in \mathcal{J}, \quad (4.26e)$$

$$(4.24b) \text{ and } (4.24d) - (4.24f). \quad (4.26f)$$

Constraint (4.26c) is still not convex, so we approximate it at the point $(p_b^{[n]}, \sigma_b^{[n]})$ as below,

$$\frac{p_b |h_b|^2}{\sigma_b^{[n]}} - \frac{\sigma_b p_b^{[n]} |h_b|^2}{\left(\sigma_b^{[n]}\right)^2} + \frac{p_b^{[n]} |h_b|^2}{\sigma_b^{[n]}} \geq \rho_b. \quad (4.27)$$

q can be obtained by exhaustive search, so constraint (4.24f) can be omitted. Thus, for a given q , the problem can be transformed to

$$\begin{aligned} & \max_{\mathbf{p}_s, p_b, \rho_b, \sigma_b} t \\ & \text{s.t. } (4.26b), (4.27), (4.26d) - (4.26f). \end{aligned} \quad (4.28)$$

Now the problem is convex and the proposed SCA algorithm is summarized in Algorithm 1, where τ is the tolerance, so that the rate region of NOMA is obtained.

Since the solution of SCA algorithm at iteration $[n]$ is also a feasible solution in iteration $[n+1]$, the objective t is non-decreasing. Therefore, the convergence of Algorithm 1 is guaranteed. In the outer loop of Algorithm 1, the computational complexity is $\mathcal{O}(N_s)$. For the inner loop, the number of iterations needed for convergence

Algorithm 1: SCA Algorithm for NOMA

```

1 Initialize:  $q = 0$ ;
2 repeat
3   Initialize:  $n \leftarrow 0, t^{[n]} \leftarrow 0, p_b^{[n]}, \sigma_b^{[n]}$ ;
4   repeat
5      $n \leftarrow n + 1$ ;
6     Solve problem (4.28) using  $p_b^{[n-1]}$  and  $\sigma_b^{[n-1]}$ , and denote the optimal
       value of the objective function as  $t^*$  and the optimal solutions as  $p_b^*,$ 
        $\sigma_b^*$ ;
7     Update  $t^{[n]} \leftarrow t^*, p_b^{[n]} \leftarrow p_b^*, \sigma_b^{[n]} \leftarrow \sigma_b^*$ ;
8   until  $|t^{[n]} - t^{[n-1]}| < \tau$ ;
9 until  $q = N_s$ ;

```

is $\mathcal{O}(\log(\tau^{-1}))$. For each iteration, the computational complexity is $\mathcal{O}[N_s]^{3.5}$ [27]. Hence, the computational complexity of Algorithm 1 is $\mathcal{O}([N_s]^{4.5} \log(\tau^{-1}))$.

4.2.3 RSMA

In the uplink RSMA, a user can split its message into multiple parts instead of sending the whole message, and then it encodes them into multiple streams and sends the superposition of these streams [36]. Therefore, the first step of RSMA design is deciding which users will split the messages. In a conventional bit communication system, if there are N users, splitting $N - 1$ users can ensure that every point on the rate region is achievable [36]. However, for the coexistence of semantic users and a bit user, will splitting semantic users contribute to a better performance? From the above discussion in Section 2.3, we know that the sentence similarity and the semantic rate are related to SINR, but splitting the semantic message will decrease the SINR. For example, we assume that a semantic user splits its message into two parts and encodes them into two streams, and the streams decoded first and decoded second are allocated power $p_{s,1,1}$ and $p_{s,1,2}$, respectively, where $p_{s,1,1} + p_{s,1,2} \leq p_{s,1}$. Thus, the

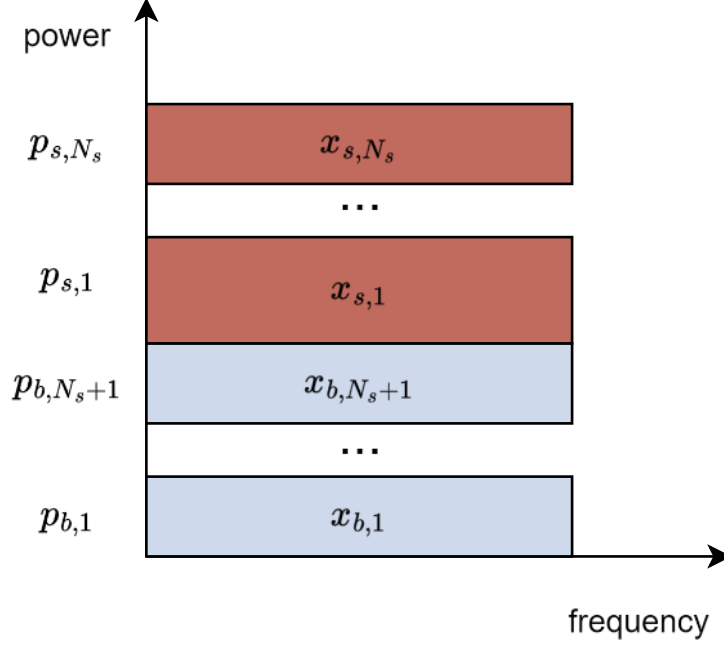


Fig. 4.5. The illustration of RSMA.

SINR of the first stream is

$$\rho_{s,1,1} = \frac{p_{s,1,1}|h_{s,1}|^2}{I_{s,1} + p_{s,1,2}|h_{s,1}|^2}, \quad (4.29)$$

where $I_{s,1}$ is the interference from other users and noise. If this semantic user does not split the message, its SINR is

$$\rho_{s,1} = \frac{p_{s,1}|h_{s,1}|^2}{I_{s,1}} \geq \rho_{s,1,1}. \quad (4.30)$$

Since this SINR decrease will deteriorate the sentence similarity, we do not split semantic users in the RSMA scheme.

After deciding on splitting the bit user only, we need to design the number of the split parts and the decoding order. Although in bit communication systems bit users usually split their messages into two parts [23, 108] because two parts can ensure that

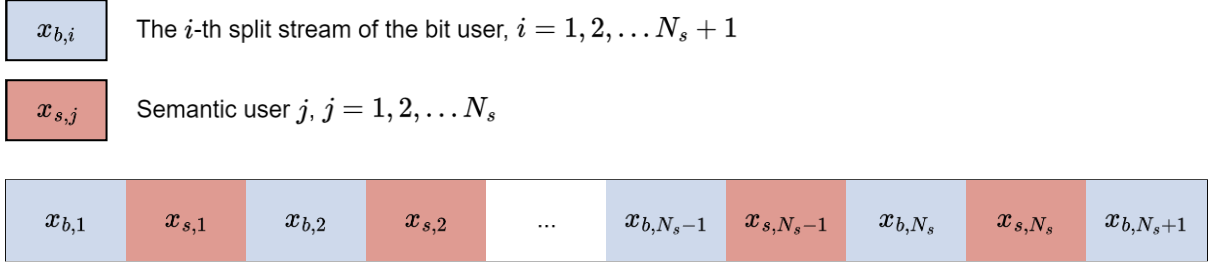


Fig. 4.6. The decoding order of RSMA.

every point in the rate region is achievable [36], for a coexistence scenario with N_s semantic users, it is more general for a bit user to split its message into $N_s + 1$ parts, which is the number of semantic users plus one, and an illustration is shown in Fig. 4.6. The semantic streams are denoted with red, and the split streams of the bit user are denoted with blue. Without the loss of generality, we assume that $x_{b,i}$ is the stream decoded i -th, i.e., $x_{b,1}$ is decoded the first and x_{b,N_s+1} is decoded the last. The semantic streams and bit streams are decoded alternatively, so that $N_s + 1$ split parts could give the most general decoding order. There are two motivations of splitting a message into $N_s + 1$ parts. First, $N_s + 1$ is the maximum number of splitting parts and it provides the most flexibility in the decoding procedure. Let us take a toy example. We assume that we aim to find the highest achievable bit rate with a given semantic rate and sentence similarity. Intuitively, the optimal solution is allocating all the power to x_{b,N_s+1} . But it is not always possible because semantic users may not tolerate much interference from the bit user, and since we do not split the semantic users, we cannot obtain any flexibility by adjusting the decoding order of the split streams of semantic users. Therefore, we can allocate part of the power to x_{b,N_s} to let part of the bit stream be decoded. If the semantic users still cannot tolerate the interference, we can allocate part of the power to x_{b,N_s-1} . In this way, at most $N_s + 1$ streams may be allocated to power. The second motivation is that adjusting the power allocation of each stream can convert $N_s + 1$ split streams to an arbitrary number of split streams between 1 and

N_s . For example, if we only allocate power to two streams and set the power of other streams as 0, then it is equivalent to having two split streams. Another advantage is that we do not need to design the decoding order. If we set the number of split streams as two, there will be $\binom{2}{N_s+1}$ decoding orders. However, if we generally set there are $N_s + 1$ split streams and perform the optimization, we do not need to consider the decoding order. If two splitting streams is the optimal solution, in the optimization result only two streams will be allocated power. For example, if there are two semantic users and the bit user splits its messages into two parts, $x_{b,1}$, $x_{b,2}$, there are three decoding orders: 1) $x_{b,1}, x_{s,1}, x_{b,2}, x_{s,2}$, 2) $x_{b,1}, x_{s,1}, x_{s,2}, x_{b,2}$ and 3) $x_{s,1}, x_{b,1}, x_{s,2}, x_{b,2}$. However, if the bit user splits its message into three parts and we set the decoding order as $x_{b,1}, x_{s,1}, x_{b,2}, x_{s,2}, x_{b,3}$, when the optimal solution is only $x_{b,1}$ and $x_{b,2}$ are allocated power, this situation is just the same as decoding order 1). Thus, we avoid considering the decoding order. Therefore, the RSMA scheme illustrated in Fig. 4.6 is the most general design for the coexistence scenario. The received signal can be represented as

$$y = \sum_{j \in \mathcal{J}} \sqrt{p_{s,j}} h_{s,j} x_{s,j} + \sum_{k \in \mathcal{K}} \sqrt{p_{b,k}} h_b x_{b,k} + n, \quad (4.31)$$

where $\mathcal{K} = \{1, 2, \dots, N_s + 1\}$, and $p_{b,k}$ is the power allocated to $x_{b,k}$. Thus, for $x_{b,k}$ the SINR is

$$\rho_{b,k} = \frac{p_{b,k} |h_b|^2}{\sigma_n^2 + \sum_{i=k}^{N_s} p_{s,i} |h_{s,i}|^2 + \sum_{l=k+1}^{N_s+1} p_{b,l} |h_b|^2}, \quad (4.32)$$

and the rate of this stream is

$$r_{b,k} = w \log_2(1 + \rho_{b,k}), \quad (4.33)$$

so the achievable rate of the bit user is

$$r_b^R = \sum_{k=1}^{N_s+1} r_{b,k}. \quad (4.34)$$

For semantic user $U_{s,j}$, the SINR is

$$\rho_{s,j} = \frac{p_{s,j}|h_{s,j}|^2}{\sigma_{s,j}}, \quad (4.35)$$

where

$$\sigma_{s,j} = \sum_{l=j+1}^{N_s+1} p_{b,l}|h_b|^2 + \sum_{i=j+1}^{N_s} p_{s,i}|h_{s,i}|^2 + \sigma_n^2, \quad \forall j \in \mathcal{J}. \quad (4.36)$$

Therefore, the achievable semantic rate is

$$S_j^R = \frac{wI}{KL} \varepsilon(\rho_{s,j}). \quad (4.37)$$

The rate region can be represented as

$$C^R \triangleq \bigcup \left\{ (S^R, R_b^R) : S^R \leq S_j^R, R_b^R \leq r_b^R, \varepsilon(\rho_{s,j}) \geq S^{th}, \right. \\ \left. j \in \mathcal{J}, p_{s,j} \leq P, \sum_{k=1}^{N_s+1} p_{b,k} \leq P \right\}. \quad (4.38)$$

For a given semantic rate S inside the rate region, a problem of maximizing the bit rate can be formulated as below:

$$\max_{\mathbf{p}_s, \mathbf{p}_b} r_b^R, \quad (4.39a)$$

$$\text{s.t. } \varepsilon\left(\frac{p_{s,j}|h_{s,j}|^2}{\sigma_{s,j}}\right) \geq S^{th}, \quad \forall j \in \mathcal{J}, \quad (4.39b)$$

$$\frac{wI}{KL} \varepsilon\left(\frac{p_{s,j}|h_{s,j}|^2}{\sigma_{s,j}}\right) \geq S, \quad \forall j \in \mathcal{J}, \quad (4.39c)$$

$$\sum_{k \in \mathcal{K}} p_{b,k} \leq P, \quad (4.39d)$$

$$p_{s,j} \leq P, \quad \forall j \in \mathcal{J}, \quad (4.39e)$$

where $\mathbf{p}_s = [p_{s,1}, \dots, p_{s,N_s}]$, $\mathbf{p}_b = [p_{b,1}, \dots, p_{b,N_s+1}]$. Since (4.39a) is not convex due

to its expression, we introduce two auxiliary variables $\boldsymbol{\rho}_b = [\rho_{b,1}, \dots, \rho_{b,N_s+1}]$, $\boldsymbol{\sigma}_b = [\sigma_{b,1}, \dots, \sigma_{b,N_s+1}]$, and we substitute (4.6) and (4.25) into (4.39b) and (4.39c), respectively, and then the problem can be transformed to

$$\max_{\mathbf{p}_s, \mathbf{p}_b, \boldsymbol{\rho}_b, \boldsymbol{\sigma}_b} \sum_{k \in \mathcal{K}} t_k, \quad (4.40a)$$

$$\text{s.t. } w \log_2(1 + \rho_{b,k}) \geq t_k, \forall k \in \mathcal{K}, \quad (4.40b)$$

$$\frac{p_{b,k} |h_b|^2}{\sigma_{b,k}} \geq \rho_{b,k}, \forall k \in \mathcal{K}, \quad (4.40c)$$

$$\sigma_{b,k} \geq \sigma_n^2 + \sum_{i=k}^{N_s} p_{s,i} |h_{s,i}|^2 + \sum_{l=k+1}^{N_s+1} p_{b,l} |h_b|^2, \forall k \in \mathcal{K}, \quad (4.40d)$$

$$\frac{p_{s,j} |h_{s,j}|^2}{\sigma_{s,j}} \geq \gamma, \forall j \in \mathcal{J}, \quad (4.40e)$$

$$\frac{p_{s,j} |h_{s,j}|^2}{\sigma_{s,j}} \geq \gamma_{sem}, \forall j \in \mathcal{J}, \quad (4.40f)$$

$$\sum_{k \in \mathcal{K}} p_{b,k} \leq P, \forall k \in \mathcal{K} \quad (4.40g)$$

$$p_{s,j} \leq P, \forall j \in \mathcal{J}. \quad (4.40h)$$

Constraint (4.40c) is still not convex, so we approximate it at the point $(p_{b,k}^{[n]}, \sigma_{b,k}^{[n]})$ as below,

$$\frac{p_{b,k} |h_b|^2}{\sigma_{b,k}^{[n]}} - \frac{\sigma_{b,k} p_{b,k}^{[n]} |h_b|^2}{\left(\sigma_{b,k}^{[n]}\right)^2} + \frac{p_{b,k}^{[n]} |h_b|^2}{\sigma_{b,k}^{[n]}} \geq \rho_{b,k}. \quad (4.41)$$

Then the problem is transformed to

$$\max_{\mathbf{p}_s, \mathbf{p}_b, \boldsymbol{\rho}_b, \boldsymbol{\sigma}_b} \sum_{k \in \mathcal{K}} t_k, \quad (4.42)$$

$$\text{s.t. } (4.40b), (4.41) \text{ and } (4.40d) - (4.40h).$$

We can obtain the rate region by solving this convex problem and the above algorithm can be summarized as Algorithm 2.

Algorithm 2 is converged because the solution of SCA algorithm at iteration $[n]$ is also

Algorithm 2: SCA Algorithm for RSMA

```
1 Initialize:  $n \leftarrow 0$ ,  $t^{[n]} \leftarrow 0$ ,  $\mathbf{p}_b^{[n]}$ ,  $\sigma_b^{[n]}$ ;  
2 repeat  
3    $n \leftarrow n + 1$ ;  
4   Solve problem (4.42) using  $\mathbf{p}_b^{[n-1]}$  and  $\sigma_b^{[n-1]}$ , and denote the optimal  
   value of the objective function as  $t^*$  and the optimal solutions as  $\mathbf{p}_b^*$ ,  $\sigma_b^*$ ;  
5   Update  $t^{[n]} \leftarrow t^*$ ,  $\mathbf{p}_b^{[n]} \leftarrow \mathbf{p}_b^*$ ,  $\sigma_b^{[n]} \leftarrow \sigma_b^*$ ;  
6 until  $|t^{[n]} - t^{[n-1]}| < \tau$ ;
```

a feasible solution in iteration $[n+1]$, which means the objective t is non-decreasing. The number of iterations needed for convergence is $\mathcal{O}(\log(\tau^{-1}))$. In each iteration, the computational complexity is $\mathcal{O}[N_s]^{3.5}$. Hence, the computational complexity of Algorithm 1 is $\mathcal{O}([N_s]^{3.5} \log(\tau^{-1}))$.

4.3 Numerical Results

In this section, we present the numerical results of the performances of different MA techniques in the coexistence scenario. The small-scale fading of the channels between the BS and all users is Rayleigh fading. The path loss is modeled as $\rho = \rho_0(1/d)^\beta$, where $\rho_0 = -30$ dB denotes the reference path loss at 1 meter, and d is the distance between the BS and the user in meters. The available bandwidth $w = 1$ MHz, and $N_0 = -140$ dBm/Hz, so that σ_n^2 over the whole available bandwidth is -80 dBm. Here we assume that the distances between the BS and the users are 30 meters and $S^{th} = 0.8$. Let d_s and d_b denote the distances between semantic users and BS and the distance between bit user and BS, respectively.

4.3.1 The rate regions for different numbers of semantic users

Firstly, let us start from a two-user case to have some intuition, which is one semantic user and one bit user and $K = 8$. We consider three cases: case 1: $d_s = 20$ meters, $d_b =$

30 meters; case 2: $d_s = 30$ meters, $d_b = 30$ meters; case 3: $d_s = 40$ meters, $d_b = 30$ meters. These three cases help us investigate how the distances affect the performances. The rate regions of three cases are shown in Fig. 4.7, and it is very different from the rate region of a two-bit-user case. For FDMA, the relation between semantic rate and bit rate is approximately linear, which is intuitive, because for both semantic user and bit user the rate increases as bandwidth increases. However, for NOMA and RSMA, the rate regions are like squares. When the semantic rate is 0, the semantic user does not transmit anything and the bit user will use all the resources to transmit, so the bit rate of NOMA and RSMA is the same as the one of FDMA. Then if the semantic user starts to transmit, r_b drops suddenly. Recall the definition of sentence similarity and the semantic rate in (2.30) and (2.31), respectively. The semantic rate is decided by sentence similarity, while that sentence similarity must be higher than S^{th} so that the semantic message can be decoded. Therefore, even if the semantic user does not require a high semantic rate, it still needs to achieve a high SINR to ensure sentence similarity. For example, for a given source of data, the semantic user $U_{s,1}$ requires $0.001 \text{ Msuts/s} \times \frac{I}{L}$, which gives that $\varepsilon(\rho_{s,1}) \geq 0.001K = 0.008$, but only if $\varepsilon(\rho_{s,1}) \geq S^{th} = 0.8$ the message can be decoded, so that $\rho_{s,1}$ must achieve a specific level. Since both NOMA and RSMA use SIC, the user decoded first will tolerate interference from the other user, so that r_b dropped suddenly to make sure that the semantic user can achieve a specific level of SINR. While FDMA allocates isolated bands to the users and they do not interfere with each other, so the relation between semantic rate and bit rate is approximately linear. Then, the bit rate of NOMA and RSMA keep the same level until the semantic rate increases to $0.1 \text{ Msuts/s} \times \frac{I}{L}$. When the semantic rate is lower than $0.1 \text{ Msuts/s} \times \frac{I}{L}$, the sentence similarity requirement is more stringent, which means as long as $\varepsilon(\rho_{s,1}) \geq S^{th}$ any semantic rate below $0.1 \text{ Msuts/s} \times \frac{I}{L}$ can be achieved. When the semantic rate is higher than $0.1 \text{ Msuts/s} \times \frac{I}{L}$, the bit rates of NOMA and RSMA drop again because the semantic rate requirement is more stringent. Due to the message splitting of RSMA, RSMA has more flexibility and

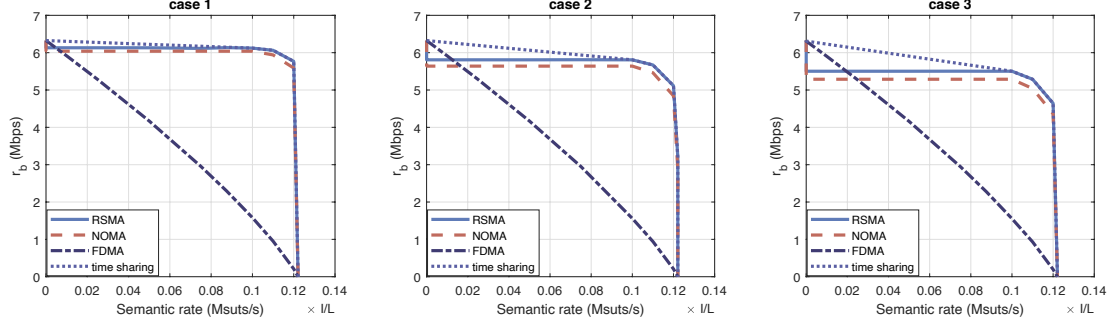


Fig. 4.7. The rate region of the coexistence of one semantic user and one bit user. X-axis denotes the semantic rate of semantic users and y-axis denotes the bit rate of the bit user. Case 1: $d_s = 20$ meters, $d_b = 30$ meters; case 2: $d_s = 30$ meters, $d_b = 30$ meters; case 3: $d_s = 40$ meters, $d_b = 30$ meters.

achieves a larger rate region. RSMA does not outperform FDMA when the semantic rate is below $0.01 \text{ Msuts/s} \times \frac{I}{L}$, because the bit user needs to sacrifice its SINR to ensure the sentence similarity of semantic users, while in FDMA the users do not interfere with each other. Nevertheless, RSMA outperforms FDMA in most situations.

Comparing three cases, we can observe that the gains of RSMA increase from case 1 to case 3, and they show that RSMA has more benefits when the bit user has better channel than the semantic user. The reason is that with a relatively better channel gain the bit user has more flexibility to split its message. Besides, we can perform time-sharing between FDMA and RSMA to obtain the largest rate region, which is shown by the dotted line in Fig. 4.7. Time-sharing is applied by dividing time with FDMA and RSMA, i.e. users use 20% time to perform FDMA and use another 80% time to perform RSMA. According to Section 4.2.3, if users use whole time to perform RSMA, the semantic user can achieve S^R and the bit user can achieve r_b^R . Therefore, when users use t_R fraction of time to perform RSMA, the semantic rate is $t_R S^R$ and the bit rate is $t_R r_b^R$. Similarly, if users perform FDMA all the time, the achievable semantic rate and bit rate are S^F and r_b^F , respectively. When they use $(1 - t_R)$ fraction of time to perform FDMA, the semantic rate and bit rate are $(1 - t_R) S^F$ and $(1 - t_R) r_b^F$, respectively. For a given S , we use exhaustive search to find the optimal t_R which gives the highest $t_R r_b^R + (1 - t_R) r_b^F$ while satisfying $t_R S^R + (1 - t_R) S^F \geq S$.

4.3.2 Comparing RSMA behaviour for bit communications and for coexistence

The main difference between RSMA scheme designs in conventional bit communications and in the coexistence scenarios is originated from the unique features of semantic communications. The effect of taking semantic features into account is intuitive if we look into the power allocation strategies. Let us still consider the two-user case to have some intuition. For the coexistence scenario, a two-user case is one bit user and one semantic user, and the bit user splits its message into two parts and encodes them into two streams. Thus, $\mathcal{K} \in \{1, 2\}$ and the power allocated to the two streams are $p_{b,1}$ and $p_{b,2}$, and we use $\alpha = \frac{p_{b,1}}{p_{b,1} + p_{b,2}}$ to denote the percentage of the power allocated to $x_{b,1}$. After we solve (4.42), we can obtain α . For a conventional two-bit-user case, only one user will split its message, and we assume that user 1 transmits $x_{1,1}$ and $x_{1,2}$ with power $p_{1,1}$ and $p_{1,2}$, respectively. User 2 does not split the message and it transmits x_2 with power p_2 . BS decodes the streams following the order of $x_{1,1}$, x_2 and $x_{1,2}$. We let h_1 and h_2 denote the channel for user 1 and user 2, respectively. For a given achievable rate of user 2, R_2 , we can find the maximum achievable rate of user 1. Therefore, the problem can be formulated as

$$\max_{\substack{p_2, p_{1,1}, p_{1,2}, \\ \rho_{1,1}, \sigma_{1,1}}} \sum_{k \in \{1, 2\}} t_k, \quad (4.43a)$$

$$\text{s.t. } w \log_2(1 + \rho_{1,k}) \geq t_k, \forall k \in \{1, 2\}, \quad (4.43b)$$

$$\frac{p_{1,1}|h_1|^2}{\sigma_{1,1}} \geq \rho_{1,1}, \quad (4.43c)$$

$$\sigma_{1,1} \geq \sigma_n^2 + p_{1,2}|h_1|^2 + p_2|h_2|^2, \quad (4.43d)$$

$$\frac{p_{1,2}|h_1|^2}{\sigma_n^2} \geq \rho_{1,2}, \quad (4.43e)$$

$$p_2|h_2|^2 \geq (2^{R_2} - 1)(\sigma_n^2 + p_{1,2}|h_2|^2), \quad (4.43f)$$

$$p_{1,1} + p_{1,2} \leq P, p_2 \leq P. \quad (4.43g)$$

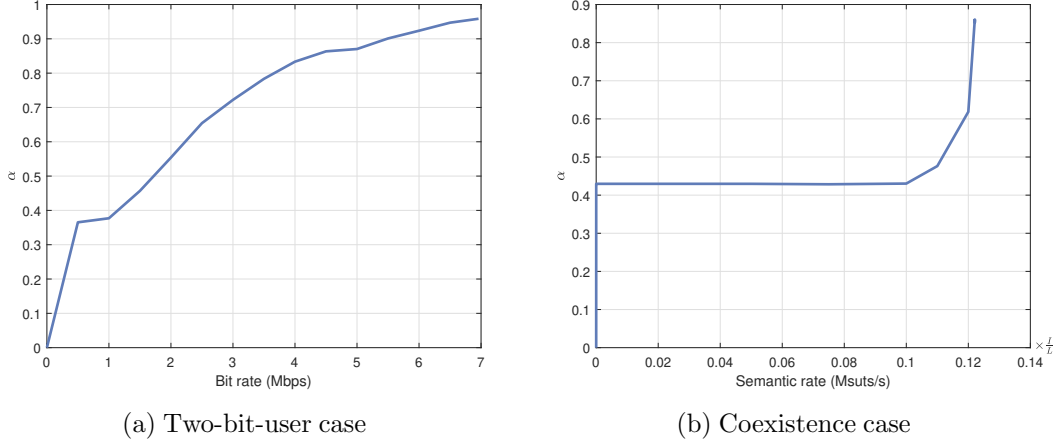


Fig. 4.8. Power allocation strategies

Then we can obtain the percentage of $p_{1,1}$, $\alpha = \frac{p_{1,1}}{p_{1,1} + p_{1,2}}$, by solving (4.43a).

Fig. 4.8 shows different power allocation strategies in two-bit-user case and coexistence case. Fig. 4.8(a) shows the relation between α and R_2 . As R_2 increases, α increases. It means that to achieve a higher R_2 , more power is allocated to $x_{1,1}$ so user 2 tolerates less interference. Fig. 4.8(b) shows the relation between α and semantic rate. When the semantic rate is 0, the semantic user does not transmit anything, so the bit user does not need to split, corresponding to $\alpha = 0$. Once the semantic user starts to transmit, it should achieve a specific SINR to ensure sentence similarity, so α achieves a value around 0.42. Then α remains unchanged until the semantic rate reaches $0.1 \text{ Msuts/s} \times \frac{I}{L}$. These changes of α are also reflected in the rate region. The bit rate drops suddenly when semantic users start to transmit, and then it maintains the same till $0.1 \text{ Msuts/s} \times \frac{I}{L}$.

4.3.3 The advantages of RSMA with multiple semantic users

The superiority of RSMA is more obvious when the number of semantic users increases. The rate regions of FDMA, NOMA and RSMA with four semantic users are shown in Fig. 4.9. The rate regions are similar to the ones in Fig. 4.7. For FDMA, the relation between the semantic rate and the bit rate is still approximately

linear, but the rate region shrinks. Since FDMA allocates isolated resources to the users, the allocated bandwidth of each user is limited for a four-semantic-user case, so it is challenging to accommodate one bit user and four semantic users with a high semantic rate. Therefore, the bit rate drops to 0 when the semantic rate is around $0.03 \text{ Msuts/s} \times \frac{I}{L}$. For NOMA and RSMA, their rate regions have similar shapes to the ones in Fig. 4.7, but the bit rates of NOMA and RSMA drop more rapidly when the semantic users start to transmit, and they drop more slowly when the semantic rate is high. Since there are four semantic users, each user should achieve a certain SINR level to make sure its message can be decoded. This causes the bit user to tolerate more interference than the one-semantic-user case, so that the achievable bit rate decreases more when the semantic users start to transmit. Then the bit rates of NOMA and RSMA remain the same as they are in the one-semantic-user case. After the semantic rate is higher than $0.1 \text{ Msuts/s} \times \frac{I}{L}$ they drop again, but more slowly compared to Fig. 4.7. In Fig. 4.7, the bit rates almost drop vertically from $0.12 \text{ Msuts/s} \times \frac{I}{L}$. Recall the approximation of sentence similarity shown in Fig. 2.5, in the high ρ regime, increasing ρ hardly improves the sentence similarity, so the semantic rate increases very slowly. Therefore, when there are only one semantic user, it can achieve a high semantic rate easily, but it is very difficult to increase the semantic rate further. Even if a small increase in semantic rate requires a huge increase in SINR. Thus, the bit rate drops almost vertically. While for a four-semantic-user case, it is not easy to achieve a high semantic rate. The semantic rate increases as the transmission power increases, so the bit user needs to tolerate more interference from semantic users and the bit rate decreases gradually from $0.1 \text{ Msuts/s} \times \frac{I}{L}$. Compared to the one-semantic-user case, the advantage of RSMA is more manifest. In the four-semantic-user case, the bit user splits its message into five parts, and it provides more flexibility to the SIC procedure compared to NOMA. For example, in NOMA the bit user may be the first to be decoded so that the semantic users can be decoded successfully. Thus, the bit rate is low because the bit user tolerates the interference from four semantic users.

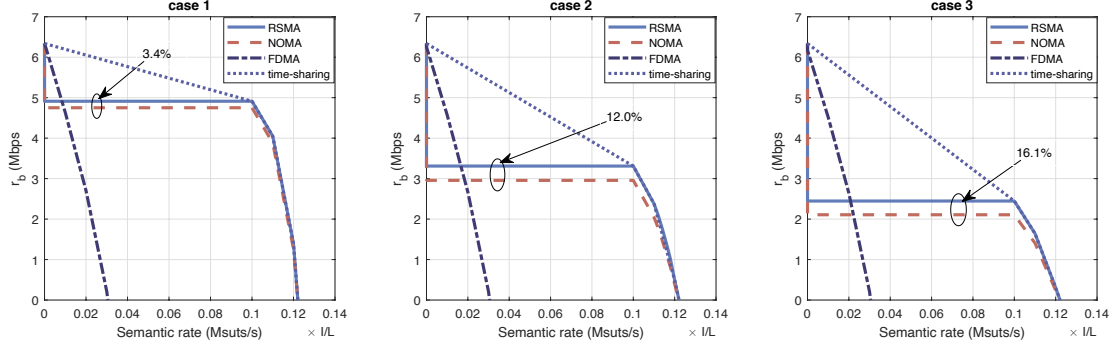


Fig. 4.9. The rate region of the coexistence of four semantic users and one bit user. X-axis denotes the semantic rate of semantic users and y-axis denotes the bit rate of the bit user. Case 1: $d_s = 20$ meters, $d_b = 30$ meters; case 2: $d_s = 30$ meters, $d_b = 30$ meters; case 3: $d_s = 40$ meters, $d_b = 30$ meters.

In RSMA, BS decodes split streams of bit user and semantic streams alternatively. Thus, the split streams do not always need to tolerate high interference, so a higher bit rate can be achieved. Similarly, the largest rate region can be achieved by performing time sharing between FDMA and RSMA, which is denoted by the dotted line.

Comparing three cases, the gains of RSMA still increases from case 1 to case 3, and they are 3.4%, 12.0% and 16.1%, respectively. They also show that RSMA has more benefits with multiple semantic users. We also observe that in most instances, splitting two streams can achieve the best performance, but sometimes the optimal result is obtained by splitting a message more than two times. Besides, we observe that when splitting two streams is the optimal solution, the decoding order is not always the same. Therefore, considering the most general case, splitting $N_s + 1$ streams, would skip the decoding order design and reduce the complexity in the optimization.

The relation between the number of semantic users and the achievable bit rate is shown in Fig. 4.10. We take case 2 as an example and still set $S^{th} = 0.8$ and $K = 8$, and we fix the semantic rate at $0.1 \text{ Msuts/s} \times \frac{1}{L}$. We do not present FDMA because the semantic users cannot always achieve $0.1 \text{ Msuts/s} \times \frac{1}{L}$ in FDMA. For both NOMA and RSMA, the relations between the number of semantic users and achievable bit rate are roughly linear. There is a constant gap between RSMA and NOMA, which shows as the number of semantic users increases the superiority of RSMA is more obvious. When there

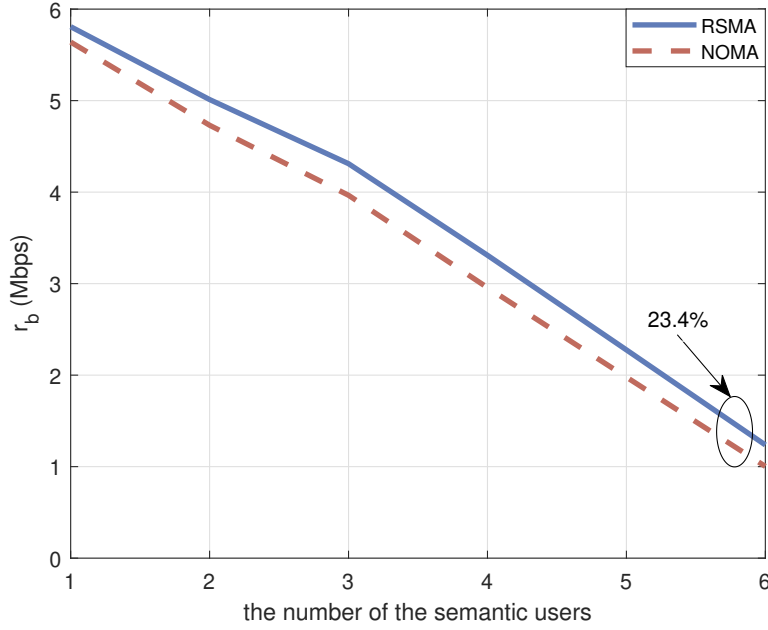


Fig. 4.10. The number of semantic users versus the achievable bit rate.

are six semantic users, RSMA has 23.4% gain comparing with NOMA. The reason is that RSMA allows the bit user to split its message into multiple parts, so it brings more flexibility during the decoding procedure and the bit user can achieve a higher bit rate.

4.3.4 The advantages of RSMA in high accuracy applications

The rate regions of case 2 with different S^{th} are shown in Fig. 4.11. $S^{th} = 0.7, 0.8, 0.9$ are denoted by green, yellow and red, respectively. RSMA, NOMA and FDMA are represented by solid line, dash line, and dash-dotted line, respectively. For FDMA, the rate regions of different S^{th} are almost the same. From Fig. 2.5, we know that the sentence similarity approaches the upper bound when SNR achieves a specific level, which means the sentence similarity hardly improves even if increasing SNR. Since the users do not interfere with each other, the SINR of each user is relatively high and their sentence similarities are very close to the upper bound. Therefore, S^{th} does not affect the resource allocation strategy. However, for NOMA and RSMA, higher S^{th} means a higher SINR of each user, so the semantic users need to use more transmission

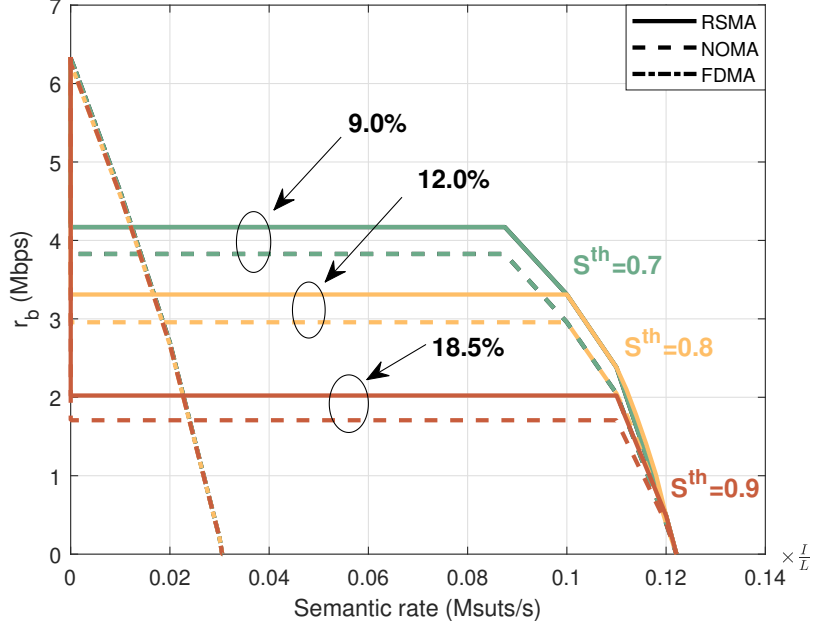


Fig. 4.11. The rate regions of $S^{th}=0.7, 0.8, 0.9$, $N_s=4$.

power. Thus, the bit user needs to tolerate more interference from others and the rate regions become smaller as S^{th} increases. Compared to NOMA, RSMA always has a larger rate region and this is benefited from message splitting. Fig. 4.11 shows that RSMA has more improvement with a higher S^{th} . When the requirement is more stringent, the flexibility of RSMA is more helpful. When $S^{th}=0.7$, the improvement of RSMA is 9.0%, and while when $S^{th}=0.9$, the improvement of RSMA is 18.5%. These results show that RSMA is more robust to a stringent sentence similarity requirement, which means it has the potential to be used in applications with high accuracy.

4.4 Summary

In this chapter, we investigate the coexistence of semantic users and a bit user in the uplink. Three MA techniques are studied for the coexistence, namely FDMA, NOMA and RSMA. Specifically, we propose a RSMA scheme for this coexistence scenario. We characterize the rate regions for FDMA, NOMA and RSMA to study their perfor-

mances. For NOMA and RSMA, to obtain the boundary points on the rate regions, we propose SCA algorithms to transform a non-convex problem into a convex problem and then solve it. We simulate the rate regions with different numbers of semantic users and various S^{th} , and we find that RSMA always outperforms NOMA but not FDMA, but we can perform time sharing to obtain the optimal performance. We summarize the main changes in RSMA design and behaviours in the coexistence scenario: 1) RSMA does not split semantic users and it splits the bit user multiple times. 2) The rate regions of the coexistence scenario and power allocation strategies are quite different from the ones of bit communications. They are caused by considering the features of semantic communications. 3) Due to introducing the semantic users, RSMA cannot always outperform OMA, but the largest rate regions can be obtained by time sharing between RSMA and OMA. Since user experiences [109], the risk of delays and reliability may not always linearly relate to the data rate, sometimes even a small increase in bit rate can bring significant improvement, so generally any use-case aiming at a higher data rate, reliability or a lower latency could be benefited from this RSMA scheme. In conclusion, the promising results show that RSMA has the potential to be applied to the coexistence of semantic users and bit users in future networks.

Chapter 5

Uplink RSMA with Hybrid ARQ

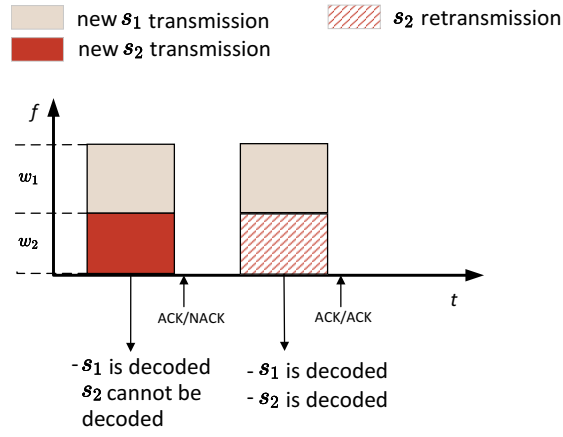
Focusing on packet transmission reliability, this chapter investigates the features of RSMA in the context of HARQ, a well-established mechanism for enhancing reliability. This chapter proposes a HARQ scheme tailored for uplink RSMA with different retransmission times for a two-user scenario and introduces a power allocation strategy for the two split streams. The results show that compared with NOMA and FDMA, RSMA outperforms them in terms of error probability and power consumption. This chapter concludes that RSMA with HARQ has the potential to improve the reliability and efficiency of wireless communication systems.

5.1 Introduction

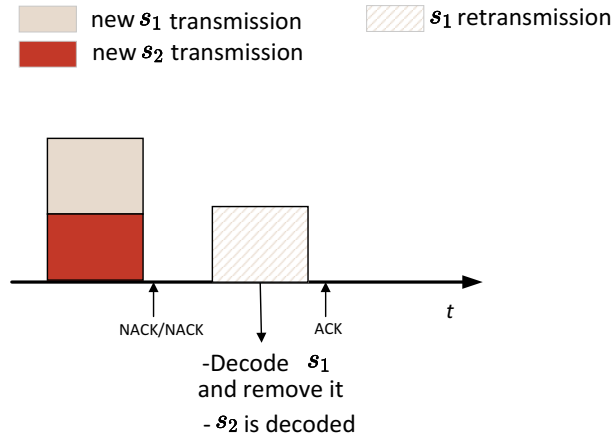
Integrating RSMA with other techniques to improve reliability is worth exploring. One critical mechanism to enhance reliability is HARQ. It also increases diversity and improves the efficiency of packet-based transmission [12]. The works on both HARQ with OMA and NOMA are reviewed in 2.4, and the literature shows that performing HARQ non-orthogonally is more efficient. However, HARQ with NOMA is doomed to face the problem that users sharing the same resources interfere with each other.

When users require the service at a high rate, even the user with a better condition may not be decoded due to interference from the other users, and it is almost impossible to decode the user with worse channels. Then in the retransmission turn, if the channel is still not good enough, an additional retransmission turn would be needed. Therefore, this could increase the latency [110].

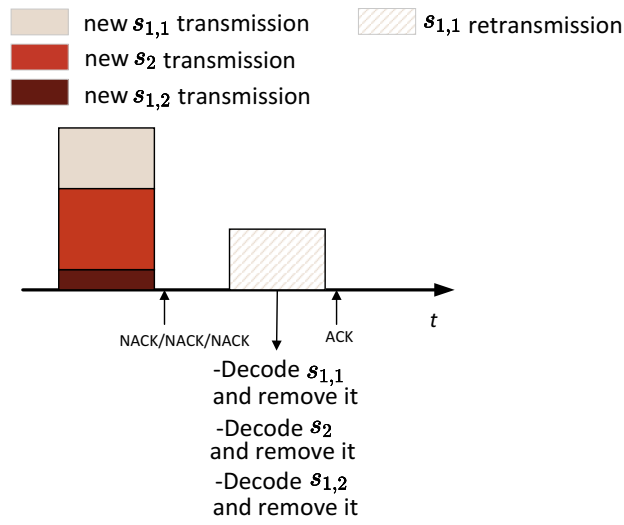
RSMA can manage interference better by bringing more streams, and it provides higher flexibility for transmission. [111] studied HARQ scheme for downlink RSMA, and showed it increased the success probability of transmission, while HARQ for uplink RSMA has not been investigated yet. Since RSMA splits its message into two parts, the number of transmitted streams increases, so the number of retransmitted streams may also increase. This increase can bring more complexity during the retransmission process. This motivates us to investigate how to design a HARQ scheme tailored for RSMA to mitigate the additional complexity. This chapter focuses on HARQ for uplink RSMA. Fig. 5.1 shows a toy example of RSMA retransmission scheme and compares it with FDMA and NOMA. Let s_1 and s_2 denote the signals transmitted by user 1 and user 2, respectively. t denotes time and f denotes frequency. Fig. 5.1(a) shows FDMA, and user 1 and user 2 occupy non-overlapping frequency bands w_1 and w_2 , respectively. If one message cannot be decoded, this message will be retransmitted and will not interfere with the other user. In Fig. 5.1(a), at first s_2 cannot be decoded and user 2 receives a NACK. Then the retransmitted s_2 and new s_1 are sent in the next time slot, and both users receive Acknowledgment (ACK). In Fig. 5.1(b), a NOMA retransmission scheme is shown. Two users can share the same time-frequency resource, and BS uses SIC to decode the two messages. We assume that user 1 has a better channel condition and s_1 will be decoded first. If s_1 fails to be decoded, neither s_2 can be decoded. But we do not need to always retransmit all the failed messages, because sometimes s_2 can be decoded directly once s_1 is decoded. Therefore, sometimes only s_1 is retransmitted, and after BS decodes s_1 successfully, s_2 is decoded. Although instead of a superposition of streams but only s_1 is transmitted, the



(a) FDMA



(b) NOMA



(c) RSMA

Fig. 5.1. The retransmission schemes of FDMA, NOMA and RSMA

retransmission run is not the same as the retransmission run of OMA. The differences are that in Fig. 5.1(a) s_2 retransmission only occupies the w_2 while in Fig. 5.1(b) the s_1 retransmission packet can occupy the whole available bandwidth, and in Fig. 5.1(a) s_2 is retransmitted along with a new s_1 while in Fig. 5.1(b) only the failed stream s_1 is retransmitted. Fig. 5.1(c) shows the retransmission scheme of RSMA. $s_{1,1}$ and $s_{1,2}$ denote the two split streams of user 1. BS needs to decode both $s_{1,1}$ and $s_{1,2}$ to obtain the message of user 1. We assume that user 1 has a better channel condition and the decoding order is $s_{1,1}$, s_2 , and $s_{1,2}$. We only split one user and set this decoding order because this setting enables the two users to achieve any point of the rate region, so it is not necessary to split both users because it would not bring further rate improvement [36]. Similarly, if $s_{1,1}$ fails to be decoded, neither s_2 and $s_{1,2}$ can be decoded. But sometimes only $s_{1,1}$ needs to be retransmitted, and s_2 and $s_{1,2}$ will be decoded once BS decodes $s_{1,1}$ successfully. In the retransmission run, sometimes only one stream is retransmitted instead of a superposition of streams, but the retransmission scheme is not exactly like FDMA for reasons similar to the ones of NOMA. Thus, compared to FDMA, RSMA has a higher spectrum efficiency thanks to full spectrum utilization and flexible interference management [112]; while compared to NOMA, RSMA can retransmit one of the split streams and has the potential to consume less energy.

There are two assumptions in this chapter. The two users require the same service, so they have the same reliability requirements. The strong user splits its message into two parts and encodes them into two streams, and the weak user encodes the whole message and transmits only one stream. The contributions of this work are summarized below:

- This chapter proposes a retransmission scheme for uplink RSMA. For a conventional HARQ scheme, the failed streams will be retransmitted. Since RSMA splits its messages into two parts, the number of transmitted streams increases. Thus, more streams may need retransmissions if they fail to be decoded. This

increase will bring more overhead and complexity during the retransmission turns. Therefore, an appropriate design HARQ scheme for RSMA is necessary. This chapter focuses on how to design a suitable HARQ scheme tailored for uplink RSMA. This scheme does not need to retransmit all the failed streams, so it will not bring additional complexity. Only one stream of the strong user and the stream of the weak user are involved in retransmission. The two streams of the strong user do not need to be decoded successively, so BS can decode these two streams and the stream of the weak user alternatively. This decoding order is optimal and it can fully exploit decoding flexibility. For the strong user, the sum rate of the two streams is the achievable rate. We assume that BS has perfect CSI, so it can decide the power allocation and decode rates of split streams. Thus, one of the streams can always be decoded successfully with a low rate by designing the power allocation appropriately, and the other stream needs to be retransmitted if the sum rate does not satisfy the desired rate. Therefore, only two of three streams involve retransmissions. To our best knowledge, HARQ with RSMA in uplink has not been studied before.

- The power allocation strategy between the two streams of the strong user is introduced in this chapter. When SIC is used, this chapter assumes that if one stream cannot be decoded, the following streams will fail. However, it is not always necessary to retransmit a sequence of failed streams. Sometimes it is possible to only retransmit one of the failed streams. Once this stream is decoded and cancelled after retransmission, BS can continue the SIC process to decode other failed streams. The power allocation between the streams of the strong users can actually decide which stream(s) needs to be retransmitted. This chapter analyzes how power allocation determines which stream(s) will be retransmitted. Naturally, different power allocation strategies bring different error probabilities of the users, and the error probabilities are given analytically in this chapter. Then, the power allocation that brings the lowest error proba-

bilities of the two users is chosen. The analysis of power allocation reveals the essence of the design of RSMA with HARQ. It explains intuitively why this HARQ scheme for RSMA has advantages with insights.

- The error probabilities and average transmission power per packet for a given rate of RSMA with HARQ are simulated by Monte Carlo and a detailed analysis is presented. The performances are simulated with both CC and IR and different retransmission times are allowed. The results show that although RSMA cannot always decrease the error probability of the strong user, it can decrease the error probability of the weak user dramatically. Since the two users require the same service, it is more critical to enable the weak user to meet the service requirements. In this way, RSMA with HARQ can support a higher rate with the same reliability requirement and retransmission times. It can achieve promising performances regardless of the latency requirement of the service. The results also show that RSMA consumes the least power compared to NOMA and FDMA, because it can mitigate the effect of interference between users. Besides, the strong user can retransmit one stream instead of two to save energy. Hence, RSMA has the potential to be applied to services requiring low power consumption.

5.2 System Model

In this section, an uplink RSMA model of a two-user case is introduced briefly. We consider the scenario that two single-antenna users transmit to a common BS with a single antenna. The two users share one time-frequency block to operate the same application with RSMA. L time retransmissions are allowed, which means a user can transmit its message up to $L + 1$ rounds, so if the message can not be successfully decoded after $L + 1$ rounds, it will be dropped. The unsuccessful copies will be buffered at BS. We assume that users do not have CSI while BS has perfect CSI.

When the users send the request for connection, BS can obtain the CSI of the users and decide who will split the message and the power allocation, and then send back this information to the users during this connection setup process and CSI is updated on each retransmission at BS. Users do not adjust transmission power because they do not have CSI, and their transmission power is normalized to 1. Let $h_k \in \mathbb{C}$ denote the channel between user k and BS, and the channel is considered as Rayleigh fading channel and fades independently, $h_k \sim \mathcal{CN}(0, \Gamma_k)$, where Γ_k is the average channel gain of user k . According to 2.1, for user k , the transmitted signal can be represented as

$$s_k = \sum_{i=1}^2 \sqrt{P_{k,i}} s_{k,i}, \quad (5.1)$$

where $s_{k,i}$ is the split stream of user k and $P_{k,i}$ is the power allocated to the stream $s_{k,i}$, which should satisfy the power constraint. The received signal at BS is

$$y = \sum_{k=1}^2 h_k s_k + n, \quad (5.2)$$

where $n \sim \mathcal{CN}(0, \sigma_n^2)$ is the additive Gaussian noise. Without loss of generality, the noise power is normalized to 1.

From 2.1, for the two-user case of RSMA, boundary points of the capacity region can be reached by one user splitting the message [23], and which user splits its message and how power is allocated can be decided by BS. We assume user 1 has a better channel condition, and it splits its message into $s_{1,1}$ and $s_{1,2}$, and $s_{1,1}$ is allocated a fraction α of the transmit power, where $\alpha \in [0, 1]$. Thus, the received signal at BS is

$$y = \sqrt{\alpha} h_1 s_{1,1} + \sqrt{1-\alpha} h_1 s_{1,2} + h_2 s_2 + n. \quad (5.3)$$

We assume that $s_{1,1}$ is the stream decoded first without loss of generality, and the decoding order is $s_{1,1}$, s_2 and $s_{1,2}$. If one stream cannot be decoded, the streams after

it are unlikely decoded. Therefore, we assume that if one stream fails, the decoding process terminates, i.e., if $s_{1,1}$ cannot be decoded, s_2 and $s_{1,2}$ will not be decoded. Thus, to obtain the whole message from user 1, BS needs to decode $s_{1,1}$, s_2 and $s_{1,2}$; while for user 2 $s_{1,1}$ and s_2 need to be decoded.

5.3 HARQ design for RSMA

This section introduces HARQ design for RSMA. Subsection 5.3.1 introduces how α affects which stream(s) will be retransmitted. α also determines the error probabilities of the two users, and the α which brings the lowest sum of error probabilities is chosen. Subsection 5.3.2 and subsection 5.3.3 present how to obtain the error probabilities with CC and IR, respectively. The HARQ schemes for FDMA and NOMA are presented in the next subsection, subsection 5.4.

5.3.1 How α Determines Which Stream Needs Retransmissions

We assume that at the BS if one stream fails, the decoding process terminates, i.e. if $s_{1,1}$ cannot be decoded, s_2 and $s_{1,2}$ will not be decoded. In the usual HARQ scheme, failed streams will be transmitted again till they are decoded or consume all the retransmission turns. Since the number of streams increases in RSMA, the complexity of HARQ will increase. However, is it necessary to retransmit all the failed streams? For example, if $s_{1,1}$ cannot be decoded, is it necessary to retransmit all the streams, $s_{1,1}$, s_2 and $s_{1,2}$ in the retransmission rounds? In fact, we do not need always to do this. In some situations, a stream can be decoded once the previous one is decoded and cancelled. In the previous example, sometimes s_2 can be decoded directly after decoding $s_{1,1}$, so we do not need to retransmit s_2 . Not only do the failed streams not always be

retransmitted together, but neither do $s_{1,1}$ and $s_{1,2}$ need to be retransmitted at the same time. Let $r_{1,1}$ and $r_{1,2}$ denote the rate of $s_{1,1}$ and $s_{1,2}$, respectively. For a given rate requirement r_1 , if $r_{1,1} + r_{1,2} \geq r_1$ holds, the message from user 1 can be decoded. In other words, if $r_{1,1} \geq r_1 - r_{1,2}$ holds, this message can be decoded. Therefore, we can only retransmit $s_{1,1}$, because it is always possible to increase $r_{1,1}$ by CC or IR during the retransmission turns and fulfils $r_{1,1} \geq r_1 - r_{1,2}$, and the maximum achievable $r_{1,2}$ is related to α . Hence, in this RSMA retransmission scheme, only $s_{1,1}$ and s_2 will be involved in retransmission. In the retransmission turn, the power allocation α will not be changed, i.e., the retransmission power of $s_{1,1}$ and $s_{1,2}$ will still be α and $1 - \alpha$, respectively. This can decrease the complexity at the receiver, otherwise, the receiver should be always informed about the new value of α . Thus, α does not need to be decided at every turn, and it can be decided when both user 1 and user 2 send new packets.

In this scheme, only $s_{1,1}$ and s_2 are involved in retransmission, and sometimes they do not need to be retransmitted together if they both fail. Channel conditions and power allocation can decide which streams will be retransmitted. Let us first consider in which situation the streams do not need to be retransmitted, and then the opposite situations would be retransmissions are needed. Let g_1 and g_2 denote the instantaneous channel gain of user 1 and user 2 in the current round, respectively. The SINR of $s_{1,1}$ is

$$\sigma_{1,1} = \frac{\alpha g_1}{1 + (1 - \alpha)g_1 + g_2}, \quad (5.4)$$

the SINR of s_2 is

$$\sigma_2 = \frac{g_2}{1 + (1 - \alpha)g_1}, \quad (5.5)$$

and the SINR of $s_{1,2}$ is

$$\sigma_{1,2} = (1 - \alpha)g_1. \quad (5.6)$$

Obviously, if there is any α that can let

$$\log_2(1+\sigma_{1,1})+\log_2(1+\sigma_{1,2})\geq r_1, \quad (5.7)$$

and

$$\log_2(1+\sigma_2)\geq r_2, \quad (5.8)$$

both hold, the three streams do not require any retransmission. From (5.7), we can obtain

$$\alpha \leq \frac{(2^{r_1}-1)(1+g_1+g_2)-g_1-g_1g_2-g_1^2}{g_1(2^{r_1}-1-g_1-g_2)} = \alpha_h, \quad (5.9)$$

and from (5.8) we can obtain

$$\alpha \geq 1 + \frac{1}{g_1} - \frac{g_2}{g_1(2^{r_2}-1)} = \alpha_l. \quad (5.10)$$

α_h and α_l may not be values between 0 and 1, and only when they are smaller than $1 + \frac{1}{g_1}$, they are mathematically meaningful, because $1+\sigma_{1,1}$, $1+\sigma_{1,2}$ and $1+\sigma_2$ should be positive. Thus, the α which can let all the streams be decoded should satisfy

$$\alpha = \forall x \in S, S = \{y: 0 \leq y \leq 1, \alpha_l \leq y \leq \alpha_h, \alpha_h \geq \alpha_l\}, \quad (5.11)$$

which means α can be any values in a set S and S is composed of the numbers between $\max(0, \alpha_l)$ and $\min(1, \alpha_h)$, as long as $\alpha_h \geq \alpha_l$.

Although α_h and α_l may not be values between 0 and 1, they can still indicate whether the desired rate pairs (r_1, r_2) are in the capacity region. Fig. 5.2 is an illustration representing the relation between the values of α_h and α_l and the location of (r_1, r_2) . The top left corner point represents $\alpha = 1$ because this point is achieved by allocating all the power to $s_{1,1}$ and decoding $s_{1,1}$ before s_2 . The bottom right corner point represents $\alpha = 0$ since this point is achieved by allocating all the power to $s_{1,2}$ and decoding s_2 before $s_{1,2}$. α between 0 and 1 indicates the points on the diagonal line. When

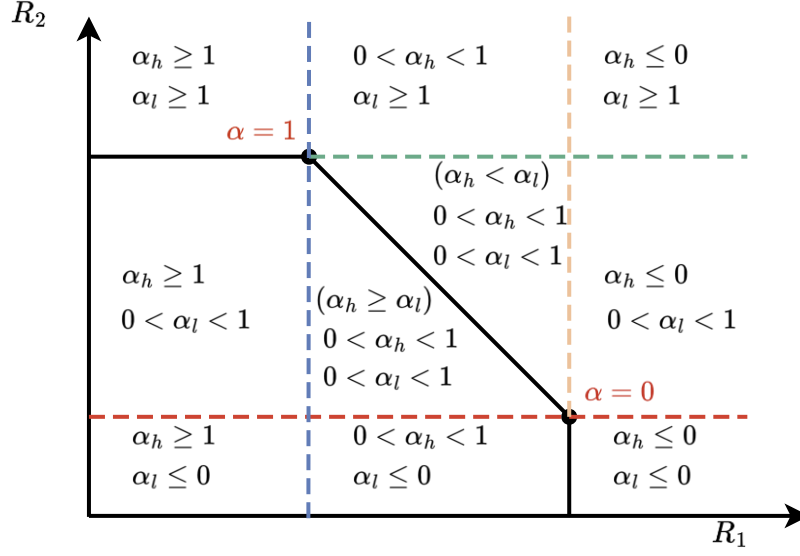


Fig. 5.2. An illustration of capacity region for two users.

$\alpha_h \leq 0$, the α satisfying (5.7) does not exist, so it represents the points on the right side of the yellow line, which means this r_1 cannot be achieved regardless of α . $\alpha_h \geq 1$ means α always satisfies (5.7), and this represents r_1 is always achievable and the corresponding points are located on the left side of the blue line. $0 < \alpha_h < 1$ represents the points between the blue line and the yellow line since r_1 is not always achievable. Similarly, $\alpha_l \leq 0$ represents the points below the red line, since any α can satisfy (5.8). While $\alpha_l \geq 1$ represents the points above the green line because no α can satisfy (5.8), and $0 < \alpha_l < 1$ represents the points between the red line and the green line.

Above analysis presents the relation between α_h , α_l and the desired rate pairs (r_1, r_2) . If an α satisfying (5.11) exists, the desired rate pair is inside the capacity region, as shown in Fig. 5.2, and no retransmission is needed. However, for some values of α_h and α_l , S in (5.11) is empty, which means α satisfying (5.11) does not exist and the rate pair is outside the capacity region. In the following situations, S is empty and retransmissions are needed.

- 1) $\alpha_h \geq 1$ and $\alpha_l \geq 1$: any α can satisfy (5.7); while no α can satisfy (5.8). In other words, s_2 needs to be retransmitted irrespective of α , while $s_{1,1}$ and $s_{1,2}$ do not need retransmission, so only s_2 needs to be retransmitted.
- 2) $0 < \alpha_h < 1$ and $\alpha_l \geq 1$: if $\alpha \leq \alpha_h$, it can satisfy (5.7); while no α can satisfy (5.8). Therefore, if we choose $\alpha \in [0, \alpha_h]$, only s_2 needs to be retransmitted. While if we choose $\alpha \in (\alpha_h, 1]$, both $s_{1,1}$ and s_2 should be retransmitted, because this α does not satisfy (5.7) and (5.8).
- 3) $\alpha_h \leq 0$ and $\alpha_l \geq 1$: no α can satisfy neither (5.7) nor (5.8), so both $s_{1,1}$ and s_2 need retransmission.
- 4) $0 < \alpha_h < 1$ and $0 < \alpha_l < 1$ ($\alpha_h < \alpha_l$): (5.7) and (5.8) cannot be satisfied at the same time. Thus, if we choose $\alpha \in [0, \alpha_h]$, only s_2 is retransmitted. While if $\alpha \in [\alpha_l, 1]$, only $s_{1,1}$ needs to be retransmitted because s_2 will be decoded after decoding $s_{1,1}$. Otherwise, $\alpha \in (\alpha_h, \alpha_l)$, both $s_{1,1}$ and s_2 are retransmitted.
- 5) $\alpha_h \leq 0$ and $0 < \alpha_l < 1$: There is α satisfying (5.8) if $\alpha \geq \alpha_l$, but no α satisfied (5.7). So if $\alpha \in [\alpha_l, 1]$, only $s_{1,1}$ needs to be retransmitted; while if $\alpha \in [0, \alpha_l)$, both $s_{1,1}$ and s_2 will be retransmitted.
- 6) : $\alpha_h \leq 0$ and $\alpha_l \leq 0$: Any α can satisfy (5.8) but no α satisfies (5.7). Thus, only $s_{1,1}$ will be retransmitted because s_2 will be decoded once $s_{1,1}$ is decoded.

These situations are also summarized in TABLE 5.1.

Apparently, for the situations that need retransmissions, the choice of α can decide which stream(s) will be retransmitted and the error probabilities after retransmissions. Since the two users operate the same service, their reliability requirements will be the same. Let $p_{1,1}$ and p_2 denote the error probability of $s_{1,1}$ and s_2 after the next retransmission turn, respectively. The α which gives the lowest $p_{1,1} + p_2$ will be chosen. Although sometimes it is possible to only retransmit one stream, considering

$\alpha_l \backslash \alpha_h$	$\alpha_h \geq 1$	$0 < \alpha_h < 1$	$\alpha_h \leq 0$
$\alpha_l \geq 1$	s_2 is retransmitted and $\alpha \in [0,1]$.	1) $\alpha \in [0, \alpha_h]$: s_2 is retransmitted. 2) $\alpha \in (\alpha_h, 1]$: both streams are retransmitted	$s_{1,1}$ and s_2 are retransmitted, and $\alpha \in [0,1]$.
$0 < \alpha_l < 1$	$\forall \alpha \in [\alpha_l, 1]$	1) $\alpha_h \geq \alpha_l$: $\forall \alpha \in [\alpha_l, \alpha_h]$. 2) $\alpha_h < \alpha_l$: if $\alpha \in [\alpha_h, 1]$, $s_{1,1}$ is retransmitted; if $\alpha \in [0, \alpha_l]$, s_2 is retransmitted; if $\alpha \in (\alpha_h, \alpha_l)$, $s_{1,1}$ and s_2 are retransmitted.	1) if $\alpha \in [\alpha_l, 1]$, $s_{1,1}$ is retransmitted. 2) if $\alpha \in [0, \alpha_l]$, $s_{1,1}$ and s_2 are retransmitted.
$\alpha_l \leq 0$	$\alpha = 0$ or $\alpha = 1$.	$\forall \alpha \in [0, \alpha_h]$	$s_{1,1}$ is retransmitted, and $\alpha \in [0,1]$.

Table 5.1: Choosing α according to α_h and α_l

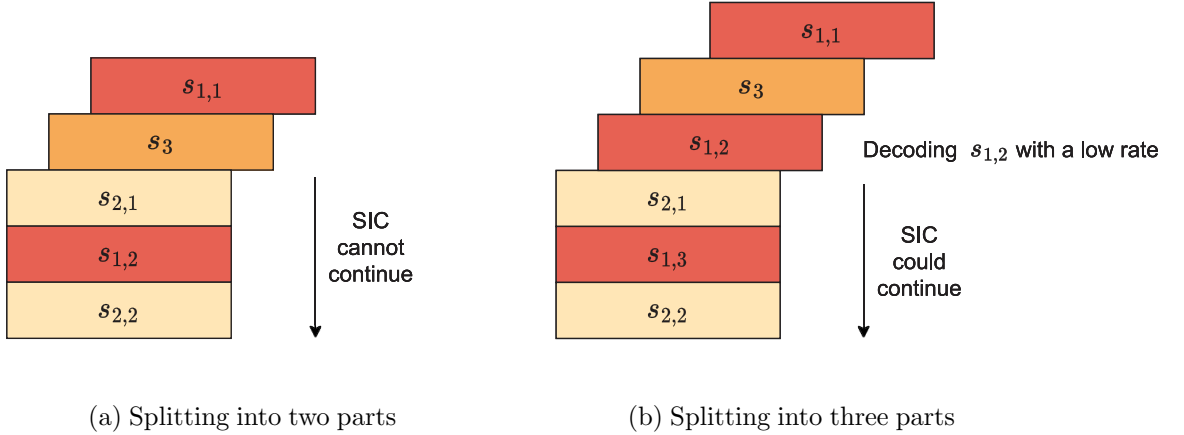


Fig. 5.3. A 3-user case

all the possibilities is for reliability reasons. When $\alpha_h \leq 0$ and $0 < \alpha_l < 1$, if only $s_{1,1}$ is retransmitted and it fails again after retransmission, s_2 still cannot be decoded; while if both streams are retransmitted, even if s_1 fails, s_2 still has some chance to be decoded. The sum of error probabilities of the former situation could be higher than the latter one. $p_{1,1}$ and p_2 in CC and IR have some differences, and the error probabilities of CC and IR will be analyzed separately.

Remark 2. *Although this work focuses on a two-user case, it can provide some insights for a general K -user case. We first discuss the most intuitive design. In a K -user case, the capacity region can be achieved by splitting $K-1$ users [36]. We assume that $k \in \mathcal{K}, \mathcal{K} = \{1, 2, \dots, K\}$ and user K does not split its message and it transmits s_K , other users split their messages into $s_{k,1}$ and $s_{k,2}$, where $k \in \mathcal{K} \setminus K$. Similar to the two-user case, we do not need to retransmit $s_{k,2}$ because the rate requirement can be fulfilled by retransmitting $s_{k,1}$. Therefore, a proper power allocation strategy would be important.*

To handle the collisions better, a user may split its message into more than two parts to have more flexibility. A simple 3-user case example can give some intuitions. For a 3-user case, the boundary points of the capacity region can be reached by splitting the messages of any two users into two parts [36], and the rate region is shown in Fig. 5.4. Therefore, from improving achievable rate perspective, splitting all the three users and more than two parts would not help. However, we want to decode $s_{k,1}$ first to avoid retransmitting $s_{k,2}$, so not all decoding orders are available in this retransmission scheme and splitting multiple parts may help. We first consider that user 1 and user 2 split their message into two parts, and the BS receives the streams $s_{1,1}$, $s_{1,2}$, $s_{2,1}$, $s_{2,2}$ and s_3 . We assume that $s_{1,1}$ and s_3 can be decoded, and $s_{1,2}$, $s_{2,1}$ and $s_{2,2}$ remain, which is shown in Fig. 5.3(a). In this case, $s_{2,1}$ needs to be retransmitted. Although splitting two parts enables the three users to achieve any point inside the rate region, splitting three parts could help from handling collisions perspective in this retransmission scheme. If user 1 splits its message into three parts, and the power fraction of $s_{1,1}$ maintains the same, the BS will receive $s_{1,1}$, $s_{1,2}$, $s_{1,3}$, $s_{2,1}$, $s_{2,2}$ and s_3 . Then $s_{1,1}$ and s_3 are decoded, and $s_{1,2}$, $s_{1,3}$, $s_{2,1}$ and $s_{2,2}$ remain. However, $s_{1,2}$ could be decoded with a low rate, and the BS may continue decoding the remaining streams, which is shown in Fig. 5.3(b). If BS still cannot continue decoding, $s_{2,1}$ will be retransmitted. This toy example shows that splitting a message into multiple parts would help handle collisions.

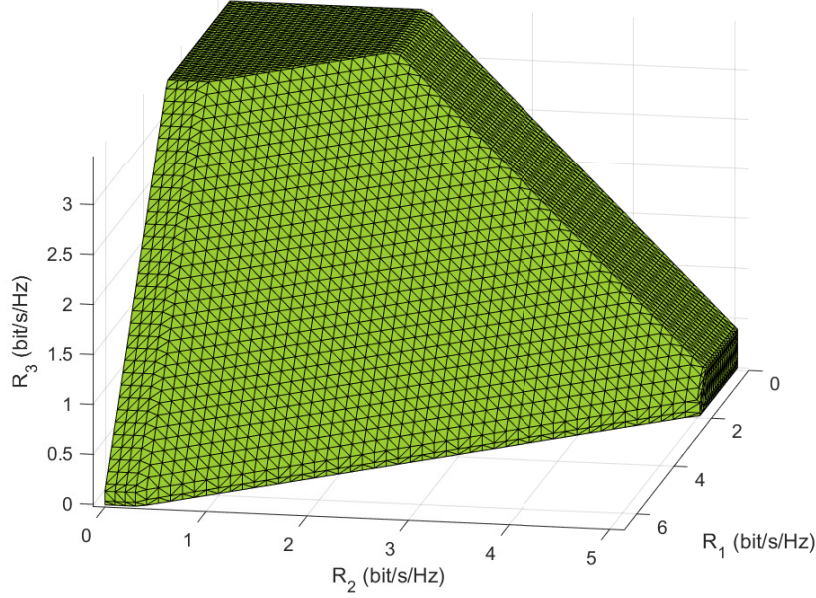


Fig. 5.4. The rate region of 3-user case

Splitting all the messages may also be helpful in this HARQ context and an illustration is shown in Fig. 5.5. s_3 may not be decoded successfully and the SIC procedure stops, which is shown in Fig. 5.5(a). If user 3 also splits its message into $s_{3,1}$ and $s_{3,2}$, $s_{3,1}$ can be decoded at a low rate and then the SIC could continue. This could handle collision better, but it will introduce more complexity. Although all users split messages may be beneficial for a HARQ scheme of RSMA, we note that it is not necessary to split all the messages to achieve the capacity region [36].

The above two toy examples show that in a general K -user case, splitting a message into multiple parts and splitting all messages could be helpful for the design of HARQ for RSMA. However, when the number of streams increases, the power allocation and decoding order should be considered carefully. Besides, increasing the number of split streams may also require increasing the buffer size, and this can lead to additional processing delays because of managing and combining more data. More split streams

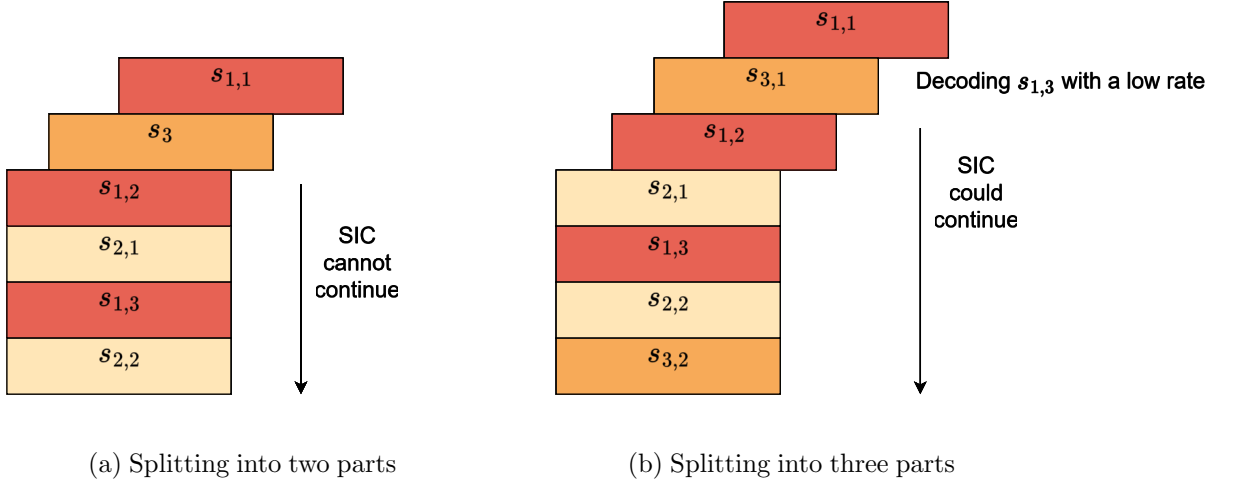


Fig. 5.5. The number of users splitting messages.

may also bring feedback delays because BS may need to send more ACK or NACK. Since BS uses SIC to decode streams, more split streams will also bring additional computation time to decide a power allocation strategy and signaling overhead. Regardless, the splitting strategy and related trade-offs are interesting topics and could be left for future work.

5.3.2 Chase Combining

When CC is used, BS can use both previous unsuccessful copies and the new packet to decode the message, and the decoding error probability after l retransmissions is given by [13],

$$p_{error}^l = \Pr \left\{ \log_2 \left(1 + \sum_{i=0}^l \text{SINR}_i \right) < R \right\}, \quad (5.12)$$

where SINR_i is the SINR of i th retransmission round, and R is the desired rate. Since SINR is related to channel gain and its Probability Density Function (PDF) is known, the error probability can be derived by integrating over the channel gain. However, for RSMA, obtaining the error probability after l retransmissions can be highly complex because of the integration in the calculation. Therefore, the choice of

α only depends on the current round and the very next retransmission round. Note that we first consider the general $0 < \alpha < 1$, and $\alpha = 0$ and $\alpha = 1$ are two special cases and they will be discussed later.

First, let us consider $p_{1,1}$ when only $s_{1,1}$ is retransmitted, which is shown in Fig. 5.1(c). If $s_{1,1}$ fails, $s_{1,1}$, s_2 and $s_{1,2}$ will all receive NACK, but only $s_{1,1}$ is retransmitted. After successfully decoding $s_{1,1}$, then s_2 and $s_{1,2}$ will be decoded, so p_2 is exactly $p_{1,1}$. Since we have obtained instantaneous channel gains g_1 and g_2 , in the following parts h_1 and h_2 denote the channels in the retransmission turn and they are all variables. BS uses the previous interfered copy of $s_{1,1}$ and retransmitted $s_{1,1}$ to jointly decode $s_{1,1}$. $p_{1,1}$ can be represented as

$$p_{1,1} = \Pr\{\log_2(1 + \sigma_{1,1} + \alpha|h_1|^2) < r_1 - \log_2(1 + \sigma_{1,2})\}, \quad (5.13)$$

and it can be rewritten as

$$p_{1,1} = \Pr\left\{|h_1|^2 < \frac{2^{r_1}}{\alpha(1 + \sigma_{1,1})} - \frac{1 + \sigma_{1,1}}{\alpha} = \gamma_{cc1,1}\right\}, \quad (5.14)$$

where $\gamma_{cc1,1}$ is a term denoting 'residual SNR' for $s_{1,1}$. Introducing this term is enlightened by [97], and it represents the needed signal power to decode the stream. Since CC and IR are discussed in separate sections, the subscript of HARQ type is omitted for simplicity, so $\gamma_{cc1,1}$ and $\gamma_{ir1,1}$ will be presented as $\gamma_{1,1}$ in this subsection and next subsection, respectively. The subscripts of HARQ type will also be omitted in other variables in the same way. The distribution of the channel gain is exponential, and the PDF is

$$f(x, \Gamma_k) = \frac{1}{\Gamma_k} e^{-\frac{x}{\Gamma_k}}, \quad (5.15)$$

so that

$$p_{1,1} = 1 - e^{-\frac{\gamma_{1,1}}{\Gamma_1}}, \quad (5.16)$$

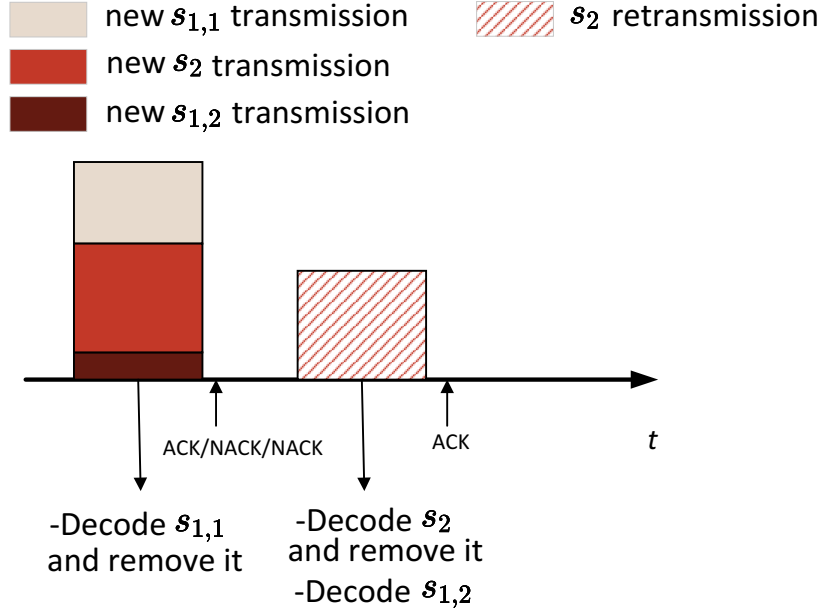


Fig. 5.6. The situation that only s_2 needs retransmission. BS fails to decode s_2 and $s_{1,2}$, but only asks user 2 to retransmit s_2 . Then, BS uses two interfered copies of s_2 to decode it, and then continues decoding $s_{1,2}$.

and $p_{1,1}$ increases as $\gamma_{1,1}$ decreases. The α which gives the lowest $p_{1,1}$ can be found by Sequential Quadratic Programming (SQP).

Then, p_2 when only s_2 is retransmitted is analyzed and an illustration is shown in Fig. 5.6. The three streams are sent to BS. BS decodes and cancels $s_{1,1}$ first, but it fails to decode s_2 , so $s_{1,2}$ also cannot be decoded. BS only asks user 2 to retransmit s_2 . Then, BS can use the unsuccessful copies of s_2 with interference and the retransmitted s_2 to decode s_2 and cancel it, and then continue decoding $s_{1,2}$. Thus,

$$\begin{aligned}
 p_2 &= \Pr\{|h_2|^2 < 2^{r_2} - 1 - \sigma_2 = \gamma_2\} \\
 &= 1 - e^{-\frac{\gamma_2}{\Gamma_2}},
 \end{aligned} \tag{5.17}$$

where γ_2 is the residual SNR for s_2 . Similar to $p_{1,1}$, p_2 decreases as γ_2 decreases. Obviously, the highest applicable α brings the lowest γ_2 , so for this situation the optimal $\alpha = \alpha_h$.

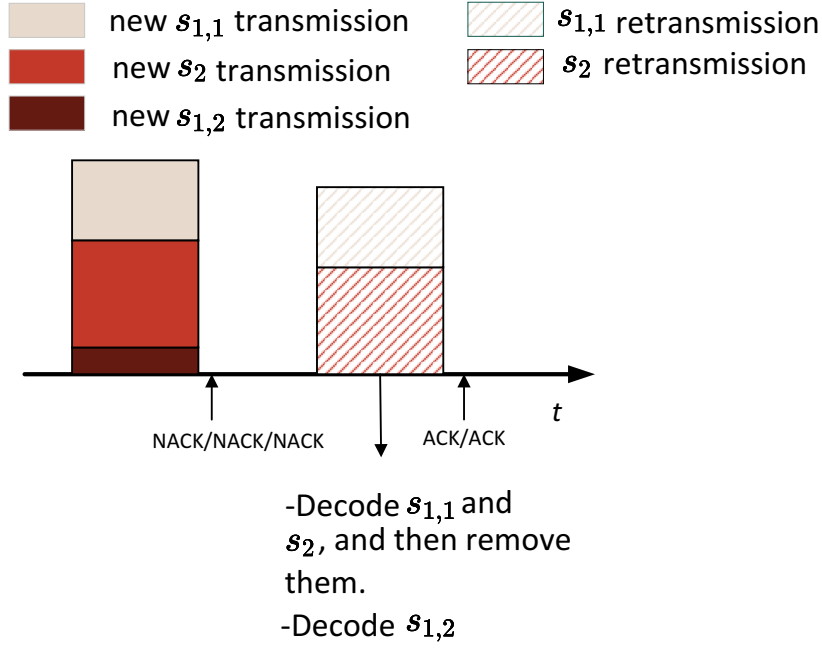


Fig. 5.7. The situation that both streams need retransmission. BS fails to decode $s_{1,1}$ and asks both user 1 and user 2 to retransmit. Then, BS uses two interfered copies of $s_{1,1}$ and s_2 to decode them and then continues decoding $s_{1,2}$.

Finally, $p_{1,1}$ and p_2 when both $s_{1,1}$ and s_2 need to be retransmitted are analyzed, and this is shown in Fig. 5.7. Although we assume that it is always possible to find a low rate to decode $s_{1,2}$, the prerequisite is that s_2 has been decoded. In the previous two situations, once $s_{1,1}$ or s_2 is decoded, the decoding process can continue, but here $s_{1,1}$ and s_2 do not depend on each other completely, so here actually $p_{1,1}$ is considered as the error probability of whole message from user 1. Thus,

$$\begin{aligned}
p_{1,1} = & \Pr \left\{ \frac{\alpha|h_1|^2}{1+|h_2|^2} < \gamma_{1,1}^{(1)}, \frac{|h_2|^2}{1+\alpha|h_1|^2} < \gamma_2^{(1)} \right\} \\
& + \Pr \left\{ \frac{|h_2|^2}{1+\alpha|h_1|^2} \geq \gamma_2^{(1)}, \alpha|h_1|^2 < \gamma_{1,1}^{(2)} \right\} \\
& + \Pr \left\{ \frac{\alpha|h_1|^2}{1+|h_2|^2} \geq \gamma_{1,1}^{(1)}, |h_2|^2 < \gamma_2^{(2)} \right\},
\end{aligned} \tag{5.18}$$

where

$$\gamma_{1,1}^{(1)} = \frac{2^{r_1}}{1+\sigma_{1,2}} - 1 - \sigma_{1,1}, \tag{5.19}$$

$$\gamma_2^{(1)} = 2^{r_2} - 1 - \frac{g_2}{1+g_1}, \quad (5.20)$$

$$\gamma_{1,1}^{(2)} = \frac{2^{r_1}}{1+\sigma_{1,2}} - 1 - \frac{\alpha g_1}{1+(1-\alpha)g_1}, \quad (5.21)$$

and

$$\gamma_2^{(2)} = 2^{r_2} - 1 - \sigma_2. \quad (5.22)$$

The error probability of s_2 can be represented as

$$\begin{aligned} p_2 = & \Pr \left\{ \frac{\alpha |h_1|^2}{1+|h_2|^2} < \gamma_{1,1}^{(1)}, \frac{|h_2|^2}{1+\alpha |h_1|^2} < \gamma_2^{(1)} \right\} \\ & + \Pr \left\{ \frac{\alpha |h_1|^2}{1+|h_2|^2} \geq \gamma_{1,1}^{(1)}, |h_2|^2 < \gamma_2^{(2)} \right\}. \end{aligned} \quad (5.23)$$

The derivation of $p_{1,1}$ and p_2 is in Appendix 1, and the results are (5.24) and (5.25), respectively,

where $c = -\frac{(1+\gamma_2^{(1)})\gamma_{1,1}^{(1)}}{\alpha(\gamma_2^{(1)}\gamma_{1,1}^{(1)}-1)}$. The α which brings the lowest $p_{1,1}+p_2$ will be chosen.

Although the above discussion is general for RSMA, $\alpha = 0$ and $\alpha = 1$ could be two special cases when considering the error probability. In the previous analysis, user 1 and user 2 will stop generating new packets in retransmission turns, because the failure of either $s_{1,1}$ or s_2 means that BS does not decode messages from user 1 and user 2 successfully. Similarly, if $\alpha = 1$ and the message of user 1 denoted by s_1 fails, both user 1 and user 2 will pause transmitting new packets. For $\alpha = 0$ and s_2 fails, the situation is the same. These two situations can be included in the above discussion. However, if $\alpha = 1$ and s_1 can be decoded, user 1 will transmit a new packet in the next turn, so user 2 will retransmit s_2 along with a new s_1 . Similarly, when $\alpha = 0$ and s_2 can be decoded, s_1 will be retransmitted with a new s_2 . In these situations, the error probability of the new message also needs to be considered, and this could affect the choice of α .

The formulations of error probabilities of these two situations are similar. We consider

$$p_{1,1} = \begin{cases} \underbrace{1 - \frac{\Gamma_1 \alpha}{\Gamma_2 \gamma_{1,1}^{(1)} + \Gamma_1 \alpha} e^{-\frac{\gamma_{1,1}^{(1)}}{\Gamma_1 \alpha}} - \frac{\Gamma_2}{\Gamma_2 + \Gamma_1 \gamma_2^{(1)} \alpha} e^{-\frac{\gamma_{1,1}^{(2)}}{\Gamma_1 \alpha} - \frac{\gamma_2^{(1)} \gamma_{1,1}^{(2)}}{\Gamma_2} - \frac{\gamma_2^{(1)}}{\Gamma_2}}}_{A_1} \\ + \underbrace{\frac{\Gamma_1}{\Gamma_1 + \Gamma_2 \gamma_{1,1}^{(1)}} e^{-\frac{\gamma_{1,1}^{(1)}}{\Gamma_1}} \left[1 - e^{-\frac{\gamma_2^{(2)}}{\Gamma_2} - \frac{\gamma_{1,1}^{(1)} \gamma_2^{(2)}}{\Gamma_1}} \right]}_{A_2}, & \text{if } \gamma_{1,1}^{(1)} \gamma_2^{(1)} \geq 1 \\ A_1 + A_2 - \frac{\Gamma_2 \gamma_{1,1}^{(1)}}{\Gamma_2 \gamma_{1,1}^{(1)} + \Gamma_1 \alpha} e^{\frac{1}{\Gamma_2} - \left(\frac{1}{\Gamma_1} + \frac{\alpha}{\Gamma_2 \gamma_{1,1}^{(1)}} \right) c} + \frac{\Gamma_2}{\Gamma_2 + \Gamma_1 \gamma_2^{(1)} \alpha} e^{-\frac{\gamma_2^{(1)}}{\Gamma_2} - \left(\frac{1}{\Gamma_1} + \frac{\gamma_2^{(1)} \alpha}{\Gamma_2} \right) c}, & \text{otherwise} \end{cases} \quad (5.24)$$

$$p_2 = \begin{cases} \underbrace{1 - \frac{\Gamma_1 \alpha}{\Gamma_2 \gamma_{1,1}^{(1)} + \Gamma_1 \alpha} e^{-\frac{\gamma_{1,1}^{(1)}}{\Gamma_1 \alpha}} - \frac{\Gamma_2}{\Gamma_2 + \Gamma_1 \gamma_2^{(1)} \alpha} e^{-\frac{\gamma_2^{(1)}}{\Gamma_2}}}_{B_1} \\ + \underbrace{\frac{\Gamma_1}{\Gamma_1 + \Gamma_2 \gamma_{1,1}^{(1)}} e^{-\frac{\gamma_{1,1}^{(1)}}{\Gamma_1}} \left[1 - e^{-\frac{\gamma_2^{(2)}}{\Gamma_2} - \frac{\gamma_{1,1}^{(1)} \gamma_2^{(2)}}{\Gamma_1}} \right]}_{B_2}, & \text{if } \gamma_{1,2}^{(1)} \gamma_2^{(1)} \geq 1 \\ B_1 + B_2 - \frac{\Gamma_2 \gamma_{1,1}^{(1)}}{\Gamma_2 \gamma_{1,1}^{(1)} + \Gamma_1 \alpha} e^{\frac{1}{\Gamma_2} - \left(\frac{1}{\Gamma_1} + \frac{\alpha}{\Gamma_2 \gamma_{1,1}^{(1)}} \right) c} + \frac{\Gamma_2}{\Gamma_2 + \Gamma_1 \gamma_2^{(1)} \alpha} e^{-\frac{\gamma_2^{(1)}}{\Gamma_2} - \left(\frac{1}{\Gamma_1} + \frac{\gamma_2^{(1)} \alpha}{\Gamma_2} \right) c}, & \text{otherwise} \end{cases} \quad (5.25)$$

$\alpha = 1$ first, and s_2 will be retransmitted with a new s_1 . For user 1,

$$p_1 = \Pr \left\{ \frac{|h_1|^2}{1 + |h_2|^2} < \gamma_1, \frac{|h_2|^2}{1 + |h_1|^2} < \gamma_2 \right\} \\ + \Pr \left\{ \frac{|h_2|^2}{1 + |h_1|^2} \geq \gamma_2, |h_1|^2 < \gamma_1 \right\}, \quad (5.26)$$

and for user 2,

$$p_2 = \Pr \left\{ \frac{|h_1|^2}{1 + |h_2|^2} < \gamma_1, \frac{|h_2|^2}{1 + |h_1|^2} < \gamma_2 \right\} \\ + \Pr \left\{ \frac{|h_1|^2}{1 + |h_2|^2} \geq \gamma_1, |h_2|^2 < \gamma_2 \right\}, \quad (5.27)$$

where

$$\gamma_1 = 2^{r_1} - 1, \quad (5.28)$$

$$p_1 = \begin{cases} \underbrace{1 - \frac{\Gamma_1}{\gamma_1 \Gamma_2 + \Gamma_1} e^{-\frac{\gamma_1}{\Gamma_1}} - \frac{\Gamma_2}{\Gamma_2 + \gamma_2 \Gamma_1} e^{-\frac{\gamma_2}{\Gamma_2} - \frac{\gamma_1}{\Gamma_1} - \frac{\gamma_1 \gamma_2}{\Gamma_2}}}_{C}, & \text{if } \gamma_1 \gamma_2 \geq 1 \\ C - \frac{\gamma_1 \Gamma_2}{\gamma_1 \Gamma_2 + \Gamma_1} e^{\frac{1}{\Gamma_2} + \frac{(1+\gamma_2)\gamma_1}{\Gamma_1(\gamma_1 \gamma_2 - 1)} + \frac{1+\gamma_2}{\Gamma_2(\gamma_1 \gamma_2 - 1)}} + \frac{\Gamma_2}{\Gamma_2 + \gamma_2 \Gamma_1} e^{-\frac{\gamma_2}{\Gamma_2} + \frac{\gamma_1(1+\gamma_2)}{\Gamma_1(\gamma_1 \gamma_2 - 1)} + \frac{\gamma_1 \gamma_2(1+\gamma_2)}{\Gamma_2(\gamma_1 \gamma_2 - 1)}}, & \text{otherwise} \end{cases} \quad (5.30)$$

$$p_2 = \begin{cases} \underbrace{1 - \frac{\Gamma_2}{\Gamma_2 + \gamma_2 \Gamma_1} e^{-\frac{\gamma_2}{\Gamma_2}} - \frac{\Gamma_1}{\Gamma_1 + \gamma_1 \Gamma_2} e^{-\frac{\gamma_1}{\Gamma_1} - \frac{\gamma_2}{\Gamma_2} - \frac{\gamma_1 \gamma_2}{\Gamma_1}}}_{D}, & \text{if } \gamma_1 \gamma_2 \geq 1 \\ D - \frac{\gamma_1 \Gamma_2}{\gamma_1 \Gamma_2 + \Gamma_1} e^{\frac{1}{\Gamma_2} + \frac{(1+\gamma_2)\gamma_1}{\Gamma_1(\gamma_1 \gamma_2 - 1)} + \frac{1+\gamma_2}{\Gamma_2(\gamma_1 \gamma_2 - 1)}} + \frac{\Gamma_2}{\Gamma_2 + \gamma_2 \Gamma_1} e^{-\frac{\gamma_2}{\Gamma_2} + \frac{\gamma_1(1+\gamma_2)}{\Gamma_1(\gamma_1 \gamma_2 - 1)} + \frac{\gamma_1 \gamma_2(1+\gamma_2)}{\Gamma_2(\gamma_1 \gamma_2 - 1)}}, & \text{otherwise} \end{cases} \quad (5.31)$$

and

$$\gamma_2 = 2^{r_2} - 1 - g_2. \quad (5.29)$$

The results are (5.30) and (5.31), and the detailed derivation is in Appendix 2.

When $\alpha = 0$, s_1 will be retransmitted with a new s_2 . The expressions are as same as (5.26) and (5.27), and the difference is the values of γ_1 and γ_2 , which are

$$\gamma_1 = 2^{r_1} - 1 - g_1, \quad (5.32)$$

and

$$\gamma_2 = 2^{r_2} - 1, \quad (5.33)$$

so by substituting (5.32) and (5.33) to (5.30) and (5.31), p_1 and p_2 can be obtained.

Then the lowest $p_1 + p_2$ can be obtained by SQP .

All the error probabilities of the situation in TABLE 5.1 can be obtained, and the α leads to the lowest error probabilities will be chosen. If the streams still cannot be decoded after one retransmission, the users will retransmit them again until the streams are decoded or they do not have more retransmission turns.

5.3.3 Incremental Redundancy

When IR is applied, the decoding error probability after l retransmissions can be represented as

$$p_{error}^l = \Pr \left\{ \sum_{i=0}^l \log_2(1 + \text{SINR}_i) < R \right\}, \quad (5.34)$$

and SINR_i is the SINR of the i th turn. α can be found by the similar method in Section 5.3.2.

If only $s_{1,1}$ is retransmitted, which is presented in Fig. 5.1(c), $p_{1,1}$ is the same as (5.16), but the 'residual SINR' has a difference. In IR,

$$\gamma_{1,1} = \frac{2^{r_1}}{\alpha(1 + \sigma_{1,1})(1 + \sigma_{1,2})} - \frac{1}{\alpha}, \quad (5.35)$$

and then impose (5.35) into (5.16), and the α which gives the lowest $\gamma_{1,1}$ will be chosen.

For the situation where only s_2 is retransmitted shown in Fig. 5.6,

$$\begin{aligned} p_2 &= \Pr \left\{ |h_2|^2 < \frac{2^{r_2}}{1 + \sigma_2} - 1 = \gamma_2 \right\} \\ &= 1 - e^{-\frac{\gamma_2}{\Gamma_2}}. \end{aligned} \quad (5.36)$$

p_2 decreases as α increases, so α_h gives the lowest p_2 .

Then, when both $s_{1,1}$ and s_2 are retransmitted in Fig. 5.7, $p_{1,2}$ and p_2 can be represented by the same expressions in (5.24) and (5.25), respectively. But

$$\gamma_{1,1}^{(1)} = \frac{2^{r_1}}{(1 + \sigma_{1,2})(1 + \sigma_{1,1})} - 1, \quad (5.37)$$

$$\gamma_2^{(1)} = \frac{2^{r_2}}{1 + \frac{g_2}{1 + g_1}} - 1, \quad (5.38)$$

$$\gamma_{1,1}^{(2)} = \frac{2^{r_1}}{(1+\sigma_{1,2})\left(1 + \frac{\alpha g_1}{1+(1-\alpha)g_1}\right)} - 1, \quad (5.39)$$

and

$$\gamma_2^{(2)} = \frac{2^{r_2}}{1+\sigma_2} - 1. \quad (5.40)$$

Substitute (5.37)-(5.40) to (5.24) and (5.25), and then the α can be found.

For the special cases when $\alpha = 0$ or $\alpha = 1$, the expressions are also the same as in (5.26) and (5.27). When $\alpha = 1$, γ_1 does not change, and

$$\gamma_2 = \frac{2^{r_2}}{1+g_2} - 1. \quad (5.41)$$

When $\alpha = 0$, γ_2 does not change, and

$$\gamma_1 = \frac{2^{r_1}}{1+g_1} - 1. \quad (5.42)$$

5.4 HARQ Scheme for FDMA and NOMA

This section will introduce the HARQ scheme for FDMA and NOMA, and we use them as baselines and compare their performances with RSMA in Section 5.5.

In FDMA, users are allocated to non-interfering resources, so each user will be allocated to an isolated bandwidth. We assume that the bandwidth is normalized to 1, and each user is allocated a fraction of the bandwidth. If CC is used, the error probability of user k after L retransmissions is

$$p_k = \Pr \left\{ w_k \log_2 \left(1 + \frac{\sum_{l=0}^L |h_k^{(l)}|^2}{w_k} \right) < r_k \right\}, \quad (5.43)$$

where w_k is the fraction of bandwidth allocated to user k , $\sum_{k=1}^2 w_k = 1$, $h_k^{(l)}$ is the

channel of user k in l th round, and r_k is the desired rate. The error probability when IR is used is

$$p_k = \Pr \left\{ \sum_{l=0}^L \left[w_k \log_2 \left(1 + \frac{|h_k^{(l)}|^2}{w_k} \right) \right] < r_k \right\}, \quad (5.44)$$

The w_k which gives the lowest $\sum_{k=1}^2 p_k$ will be chosen, and it can be computed by Monte Carlo simulation. Compared to RSMA, the retransmissions are independent in FDMA, i.e., whether a user will retransmit does not depend on other users, and the users retransmit whole streams.

NOMA can be seen as a subset of RSMA since it equals that $\alpha = 0$ or $\alpha = 1$. The decoding order is decided by the channel gain, which means if user 1 has better channel conditions, it will be decoded first ($\alpha = 1$), otherwise, user 2 will be decoded first ($\alpha = 0$). For CC, the error probability of user 1 is

$$p_1 = \Pr \left\{ \log_2 \left(1 + \sum_{l=1}^L \text{SINR}_1^{(l)} \right) < r_1 \right\}, \quad (5.45)$$

where

$$\text{SINR}_1^{(l)} = \begin{cases} \frac{|h_1|^2}{1 + |h_2|^2}, & \text{if } \alpha = 1 \\ |h_1|^2, & \text{if } \alpha = 0. \end{cases} \quad (5.46)$$

The error probability of user 2 is

$$p_2 = \Pr \left\{ \log_2 \left(1 + \sum_{l=1}^L \frac{|h_2^{(l)}|^2}{1 + (1 - \alpha)|h_1^l|^2} \right) < r_2 \right\}. \quad (5.47)$$

For IR, similarly, the error probability of user 1 is

$$p_1 = \Pr \left\{ \sum_{l=1}^L \left[\log_2 \left(1 + \text{SINR}_1^{(l)} \right) \right] < r_1 \right\}, \quad (5.48)$$

where $\text{SINR}_1^{(l)}$ is (5.46), and error probability of user 2 is

$$p_2 = \Pr \left\{ \sum_{l=1}^L \left[\log_2 \left(1 + \frac{|h_2^{(l)}|^2}{1 + (1 - \alpha)|h_1^l|^2} \right) \right] < r_2 \right\}. \quad (5.49)$$

NOMA can be seen as a special case of RSMA, but it only retransmits whole streams, so it is not as flexible as RSMA.

5.5 Numerical Results

In this section, the error probabilities and transmission power per packet of RSMA with HARQ will be presented, and they are compared with FDMA and NOMA. For all the simulation results, the average channel gains, Γ_1 and Γ_2 , are set to 20 dB and 15 dB, respectively. Since we assume that user 1 and user 2 require the same service, they have the same reliability requirement and the desired rate, and this rate is the x-axis in the figures.

In Fig. 5.8 and Fig. 5.9, two retransmissions are allowed with CC, i.e. for low latency service, URLLC. In Fig. 5.8, the error probabilities for different rates are presented, and they are obtained by Monte Carlo simulation. The solid lines denote user 1, and the dashed lines marked by triangles denote user 2. Generally, user 1 has a lower error probability than user 2 due to the better channel condition, and RSMA has the lowest error probability for user 2. Actually, the error probability of user 2 is more critical, because it also needs to meet the reliability requirement. That both users can have a low error probability is a main advantage of RSMA. The reason is that RSMA can exploit the resources better than FDMA, and it is more flexible than NOMA. In NOMA, if user 1 cannot be decoded, user 2 will also fail; while in RSMA, part of the interference of user 1 can be cancelled and user 2 still has the possibility to be decoded. Thus, RSMA has better performance for user 2. For example, at the initial

round, the magnitudes of channel gain of user 1 and user 2 are 35.6954 and 26.3955, respectively, and the aim rate is 3 bits/s/Hz. BS is aware of CSI and decide to only ask user 1 to retransmit. In the next round, the magnitude of channel gain of user 1 is 1.8654, and BS asks user 1 to transmit again. In the last retransmission round, user 1 still does not have good channel and the magnitude of channel gain is 2.7076. In this case, user 1 can only achieve $r_1 = \log_2\left(1 + \frac{35.6954}{1+26.3955} + 1.8654 + 2.7076\right) = 2.8996$ bits/s/Hz, so it cannot be decoded and user 2 also fails. While for RSMA, if α is set to 0.9224 and retransmit $s_{1,1}$, user 1 will be decoded after one retransmission with rate $r_1 = \log_2\left(1 + \frac{35.6954\alpha}{1+35.6954(1-\alpha)+26.3955} + 1.8654\alpha\right) + \log_2(1 + 35.6954(1-\alpha)) = 3.8450$ bits/s/Hz, and user 2 with rate $r_2 = \log_2\left(1 + \frac{26.3955}{1+35.6954(1-\alpha)}\right) = 3.0004$ bits/s/Hz. Thus, the error probability of user 2 decreases. Fig. 5.9 shows the average transmission power per packet. RSMA consumes the least energy. One reason is that it has the largest capacity region than FDMA and NOMA without time-sharing so it is more possible to achieve a given rate. For example, the magnitudes of channel gain of user 1 and user 2 are 162.4555 and 85.2316, respectively, and the aim rate is 2 bits/s/Hz. NOMA needs to retransmit s_1 , while for RSMA α can be set to 0.9 to avoid retransmission. The other reason is that when $s_{1,1}$ is retransmitted only α fraction of power is used. NOMA consumes more power than FDMA, because in NOMA s_1 is decoded with interference, so s_1 is retransmitted a bit more frequently than FDMA, which is free from interference from others.

In Fig. 5.10 and Fig. 5.11, IR is used and two retransmissions are allowed. Fig. 5.10 presents the error probabilities of user 1 and user 2. RSMA with HARQ still has the best performance and the two users can all obtain low error probabilities. The trends of the curves are quite similar to the results in Fig. 5.8, but FDMA outperforms NOMA. FDMA benefits more from IR since the achievable rate is the sum of the rate in each turn, which mitigates the effect of limited bandwidth. While for NOMA the retransmission with interference may not contribute to the achievable rate as much as the non-interfering retransmission, so the error probability increases. However, since

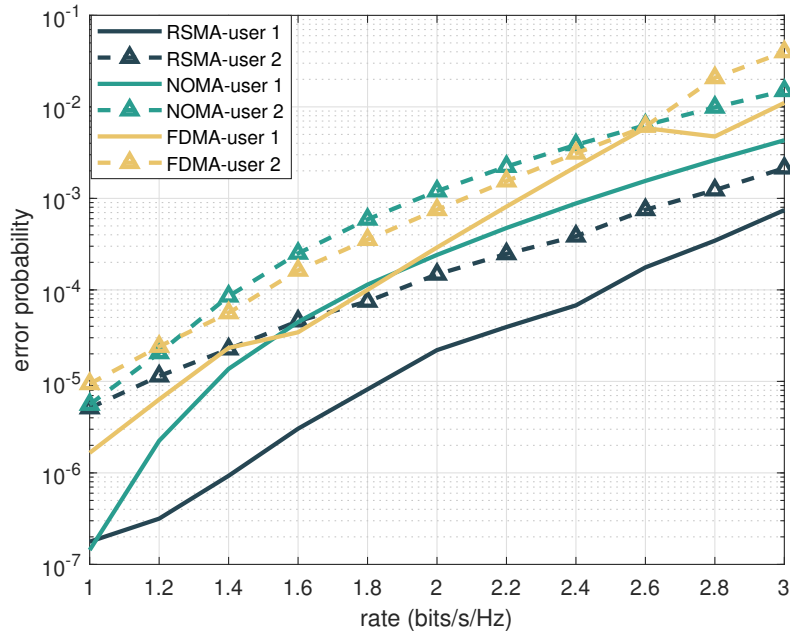


Fig. 5.8. Error probabilities of two users in CC. The average channel gains for user 1 and user 2 are 20 dB and 15 dB, respectively. Retransmission times $L = 2$.

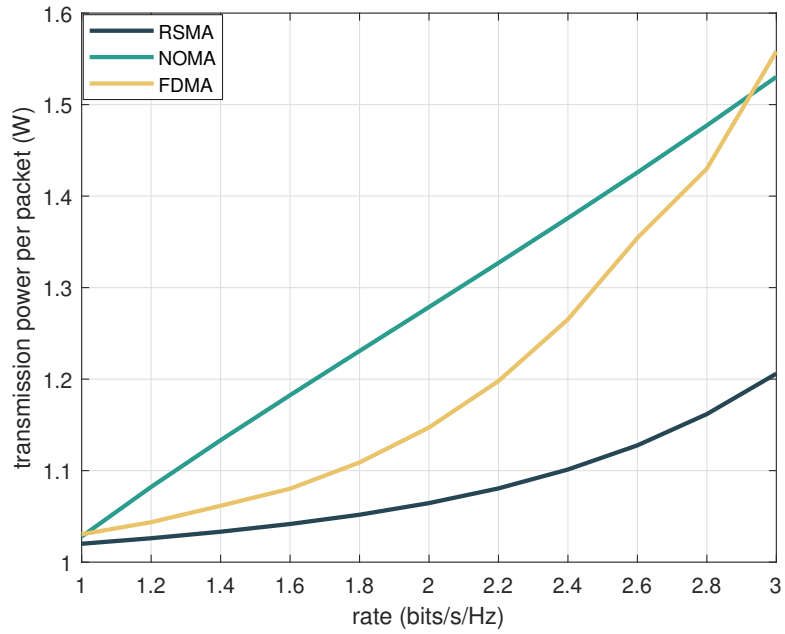


Fig. 5.9. Average power consumption of each packet in CC with 2 retransmissions.

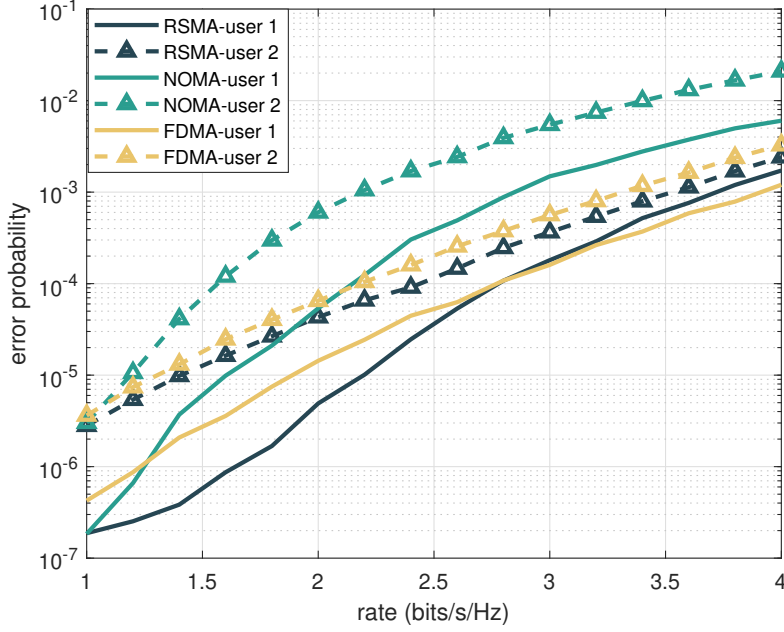


Fig. 5.10. Error probabilities of two users in IR. The average channel gains for user 1 and user 2 are 20 dB and 15 dB, respectively. Retransmission times $L = 2$.

the decoding order of RSMA is shuffled, the negative impact of interference is mitigated to some extent. For example, in RSMA, when α is between 0 and 1 and either $s_{1,1}$ or s_2 needs to be retransmitted, the stream will be retransmitted alone without interference. But the decoding of $s_{1,2}$ depends on s_2 , and this is the reason the error probability of user 1 is close to user 2 as the rate increases. While in NOMA, when s_1 is decoded before s_2 , if s_1 can be decoded and s_2 fails, s_2 will be retransmitted with another new s_1 , and s_2 could fail again due to the interference of this new stream. Fig. 5.11 shows the average power consumption of each packet. Similar to the trends in Fig. 5.9, RSMA consumes the least power. NOMA consumes more power than FDMA when the rate is lower than around 2.7 (bit/s/Hz), because NOMA could need more retransmissions than FDMA. However, NOMA consumes less power when the rate is higher than 2.7 (bit/s/Hz) since FDMA is limited by isolated resources.

Fig. 5.12 and Fig. 5.13 present the error probabilities and average power consumption per packet with CC, respectively, and 4 retransmissions are allowed. This set can be seen as for the service which is not very sensitive to the latency. In Fig. 5.12, for

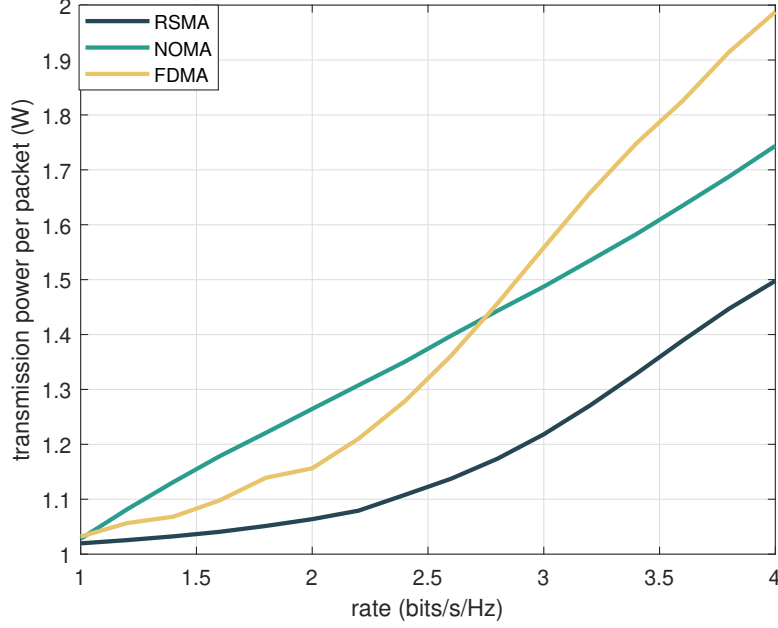


Fig. 5.11. Average power consumption of each packet in IR with 2 retransmissions.

both two users, RSMA has the best performance, NOMA has the second best, and the error probabilities in FDMA are the highest. FDMA curves are not smooth because the objective is minimizing the error probabilities, so when the rate is higher than 4.5 (bit/s/Hz) it allocates more resources to user 2 to minimize the error probabilities. FDMA with CC is not quite efficient when the rate is high due to the limited bandwidth while allocating resources in a non-orthogonal manner can exploit the resources better. Although NOMA has a much better performance than FDMA, for user 2 there is still room to improve. A simple example can give intuition. We assume that s_1 is decoded before s_2 . Decoding s_1 at a relatively high rate is not always possible due to the interference from s_2 , and decoding s_2 actually depends on whether decoding s_1 successfully, so the situation that both users cannot be decoded may happen. In RSMA, $s_{1,1}$ can be decoded and cancelled at a relatively low rate, and this increases the probability of decoding s_2 successfully. Besides, the two users will stop generating new packets if either $s_{1,1}$ or s_2 fails. Thus, if only one stream needs to be retransmitted, it will be retransmitted without interference. When 4 retransmissions are allowed, these non-interfering copies can contribute more to achieve a higher rate. The average

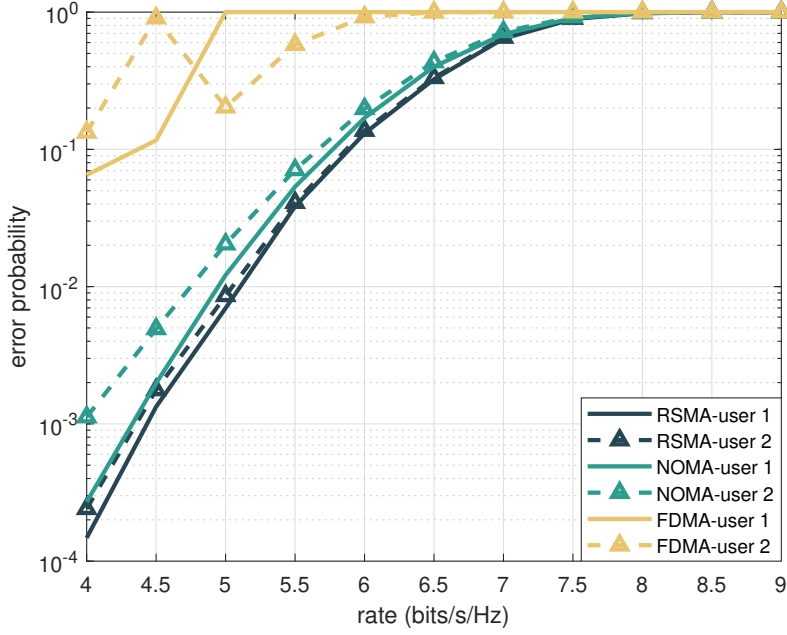


Fig. 5.12. Error probabilities of two users in CC. The average channel gains for user 1 and user 2 are 20 dB and 15 dB, respectively. Retransmission times $L = 4$.

power consumption is shown in Fig. 5.13. FDMA consumes the most power per packet among the three schemes. In FDMA, each user occupies a limited bandwidth, so the frequency of retransmission increases. RSMA and NOMA consume less power since they do not retransmit failed streams all the time. RSMA consumes the least power when the rate is lower than 8 (bit/s/Hz). When the rate is higher than 8 (bit/s/Hz), the power consumption of RSMA and NOMA are almost the same, because RSMA tends to allocate more power to $s_{1,1}$ so that s_2 can experience less interference. Besides, the frequency of retransmitting $s_{1,1}$ and s_2 together increases. Although RSMA may boil down to NOMA, it still has an advantage in power consumption, because at 5 (bit/s/Hz) the error probabilities of NOMA and RSMA are around 10^{-2} , and services may not desire an error probability which is higher than 10^{-2} .

The performances when IR is applied with 4 retransmissions are shown in Fig. 5.14 and Fig. 5.15. Compared to HARQ-CC, the achievable rate of HARQ-IR is much higher when 4 retransmissions are allowed since the logarithmic function is a concave function. The error probabilities of the two users are shown in Fig. 5.14. According

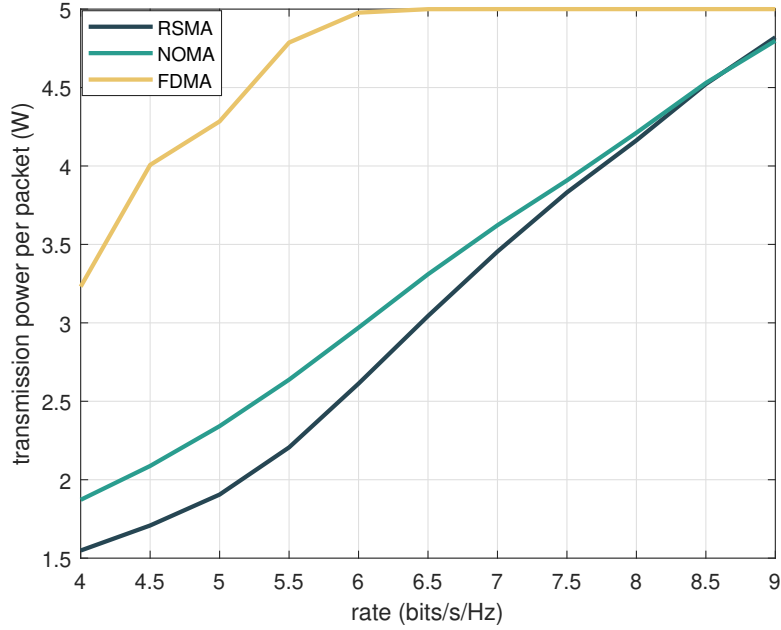


Fig. 5.13. Average power consumption of each packet in CC with 4 retransmissions.

to Fig. 5.14, NOMA has the highest error probabilities for both two users. While for user 1 FDMA has a lower error probability than RSMA, RSMA has a lower error probability for user 2 than FDMA. When a high rate is required and the latency requirement is not strict, it would be better to let isolated resources be dedicated to the users, because the user with good channel conditions cannot tolerate that much interference. The copies with interference would also not help much, which is also shown in Fig. 5.10. Hence, FDMA has lower error probabilities than NOMA though the users can only occupy part of the resources. Although users can interfere with each other in RSMA, in the retransmission rounds they do not always interfere with each other. Similar to RSMA with CC, that either $s_{1,1}$ or s_2 fails will cause both users to pause generating new packets, so $s_{1,1}$ or s_2 will not be retransmitted with the interference from a new message. In this way, the two users can share the resources but not always interfere with each other. Fig. 5.15 shows the average transmission power for each packet. FDMA consumes the most power, NOMA follows, and RSMA is the lowest. For NOMA and RSMA, it is not necessary to transmit all the failed streams, so they consume less power than FDMA. RSMA uses the least power because it can

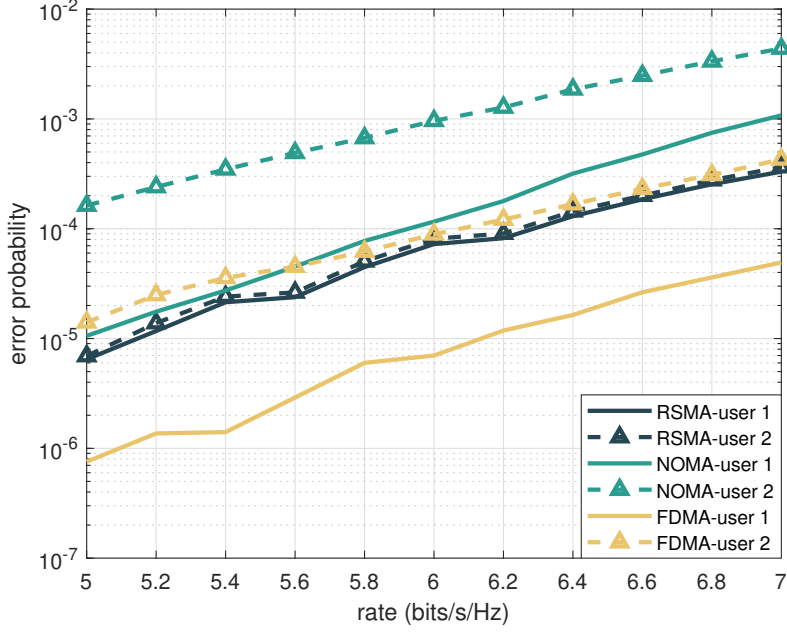


Fig. 5.14. Error probabilities of two users in IR. The average channel gains for user 1 and user 2 are 20 dB and 15 dB, respectively. Retransmission times $L=4$.

mitigate the effect of interference to some extent, so fewer retransmissions are needed compared to NOMA.

According to the simulation results, RSMA with HARQ has the potential to improve the achievable rate for users who require the same service. Two users can have relatively low error probabilities with both CC and IR and different retransmission times. The simulation results also show that RSMA with HARQ has the lowest average power consumption per packet, which means it is suitable for the service of low-power devices.

5.6 Summary

In this chapter, a retransmission scheme tailored for uplink RSMA is proposed. This scheme does not need to retransmit all the failed streams, so it avoids bringing additional complexity. This chapter focuses on the scenario that two users requiring the

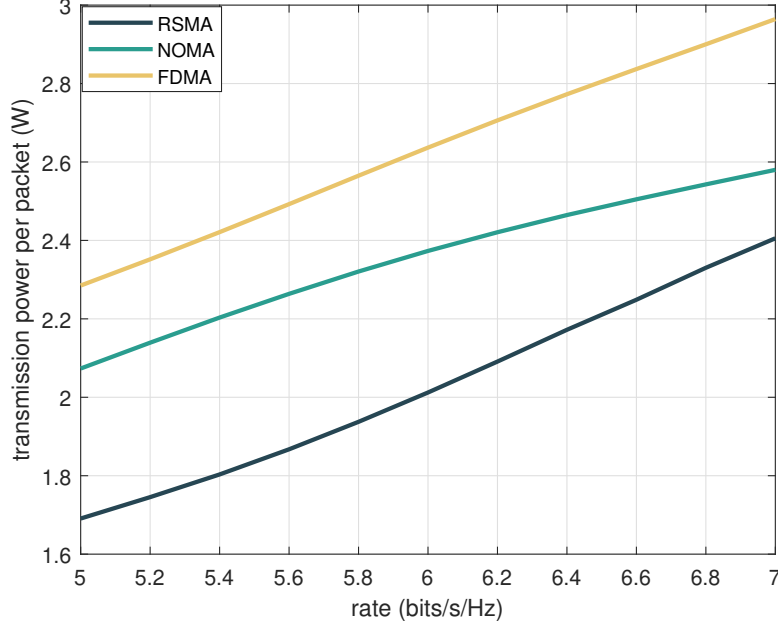


Fig. 5.15. Average power consumption of each packet in IR with 4 retransmissions.

same service share a common BS. The decoding order at BS is set to $s_{1,1}$, s_2 , and $s_{1,2}$, which is the optimal order to exploit the flexibility of RSMA. It is always possible for $s_{1,2}$ to be decoded with a low rate, so only $s_{1,1}$ and s_2 are involved with retransmitting. Power allocation between $s_{1,1}$ and $s_{1,2}$ is important and it decides which stream(s) will be retransmitted. Error probabilities for different power allocation strategies are given, and the α which gives the lowest sum of the error probabilities is chosen.

The simulation results present the error probabilities and the average power consumption per packet of RSMA with CC and IR, and they are compared with NOMA and FDMA. The results show that RSMA can let both users have a low error probability while consuming less energy with different retransmission times. This shows that RSMA with HARQ could be applied to services that need low latency and low power consumption. RSMA with retransmission can enhance diversity and improve efficiency, so it could be a promising scheme for future communication networks.

Chapter 6

Conclusion

6.1 Summary of Thesis Achievements

In this thesis, we investigate the application of RSMA to network slicing and the coexistence of bit communications and semantic communications. A HARQ scheme tailored for RSMA is investigated to enhance reliability. Simulation results and analysis are provided to evaluate the performance of RSMA.

In Chapter 3, we investigate eMBB and URLLC coexistence and eMBB and mMTC coexistence with OMA, NOMA and RSMA. Deciding which user splits the messages and power allocations are significant for RSMA scheme designs. The RSMA schemes are designed for two coexistence scenarios based on the different requirements of eMBB, URLLC and mMTC. For eMBB and URLLC coexistence, numerical results show the achievable rate pairs of eMBB and URLLC in different channel conditions, and how the splitting power factor impacts the performances. For eMBB and mMTC coexistence, a novel RSMA scheme is proposed. The pairs of eMBB rate and mMTC arrival rate are simulated, and the impact of the splitting power factor is also analysed. The results show that RSMA can always outperform NOMA counterpart and

can achieve the pairs that OMA cannot.

In Chapter 4, we investigate the coexistence of semantic users and a bit user in the uplink. Three MA schemes are considered: FDMA, NOMA and RSMA, and the rate regions of these three schemes are characterized. The main novelty is that we propose a RSMA scheme for this coexistence scenario, and analyses the differences between RSMA designs in coexistence and bit communications. To evaluate the performances, a SCA algorithm is proposed to obtain the rate region. Numerical results show that RSMA has the potential to be applied to the coexistence scenario of semantic users and bit users.

In Chapter 5, a HARQ scheme tailored for uplink RSMA is proposed. In this scheme, the users do not need to retransmit all the failed streams, so it will not bring additional complexity. This chapter focuses on a two-user case, and the power allocation strategies are designed to minimize the error probabilities of users, and the expressions of error probabilities are provided. Simulation results show that RSMA with HARQ outperforms NOMA and FDMA in terms of error probabilities and power consumption. This shows that RSMA has the potential to be applied to the scenario requiring high reliability.

In conclusion, RSMA is a promising MA technique in terms of spectrum efficiency and interference management. As a non-orthogonal MA technique, RSMA allows users to share the resources, which brings higher spectrum efficiency than OMA, so it can better handle the scenarios with more users. Compared to NOMA, RSMA has a more flexible interference management strategy. In uplink, this flexibility is mainly reflected in the flexible decoding order, and this is especially useful for the system aiming at user fairness and users with various requirements. Chapter 3, Chapter 4 and Chapter 5 show that how this flexibility helps to improve the performance in some scenarios of future networks. Although flexible decoding order and resource allocation enables RSMA to achieve promising performances, flexibility also means that decoding order

and resource allocation should be considered carefully. Overall, RSMA shows its potential in future networks to handle the requirements of various services and users, but there are still topics to be explored and some of them are presented in the next part.

6.2 Future Work

Some potential future research directions are listed as follows:

1. **More general setups of RSMA for network slicing:** In Chapter 3, for eMBB and URLLC coexistence scenario, we do not consider finite blocklength, because in this work we consider that URLLC does not have instantaneous CSI, such that the factor which causes packet error is the randomness of the channel coefficient (SNR) instead of the noise instance at a given SNR [113]. Therefore, we are working with the asymptotic rate. The scenario related to finite blocklength could be explored in future work. A more general n_U -URLLC-user scenario would also be an interesting direction and the extension of Remark 1 about decoding order could be studied.
2. **Multiple bit users and various semantic information sources in coexistence scenario:** In Chapter 4, a scenario with one bit user and multiple semantic users is investigated. How to accommodate multiple bit users and multiple semantic users can be explored in future. Especially, the number of split streams and the decoding order are interesting questions to explore. Besides, Chapter 4 focuses on the semantic communication for text transmission, DeepSC, it would be also interesting to explore how to design RSMA in the coexistence scenario semantic communications for other information sources, i.e. images, videos and speeches. Besides, since DeepSC is an end-to-end framework, Chapter 4 only considers SISO, but exploring a coexistence scenario with multiple antennas and how to design RSMA would be interesting topics.

3. RSMA with HARQ in multiple user cases: Although Chapter 5 focuses on the two-user case, it also provides insights for a multiple-user case. The decoding order and resource allocation need to be considered carefully. In some cases, splitting a message into multiple parts may avoid the collision. It is interesting to explore when it would be helpful to split a message into multiple parts, how to design HARQ for RSMA in such a case, and to what extent collisions could be avoided. These topics can be left as future work. This chapter considers that users keep transmitting packets, and investigating the scenario that users send packets with a specific probability could be considered in future. This direction can also be integrated with random access to study the case that numerous users are connected to one BS. Besides, this work aims to study the fundamental limits of the systems and assumes perfect SIC, while how imperfect SIC will impact the performance is a critical issue and it can be an interesting topic.

References

- [1] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufveson, “6G wireless systems: Vision, requirements, challenges, insights, and opportunities,” *Proceedings of the IEEE*, vol. 109, no. 7, pp. 1166–1199, 2021.
- [2] H. Zhang, N. Liu, X. Chu, K. Long, A. Aghvami, and V. C. M. Leung, “Network slicing based 5G and future mobile networks: Mobility, resource management, and challenges,” *IEEE Communications Magazine*, vol. 55, no. 8, pp. 138–145, 2017.
- [3] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, “Network slicing in 5G: Survey and challenges,” *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94–100, 2017.
- [4] P. Zhang, W. Xu, H. Gao, K. Niu, X. Xu, X. Qin, C. Yuan, Z. Qin, H. Zhao, J. Wei, and F. Zhang, “Toward wisdom-evolutionary and primitive-concise 6G: A new paradigm of semantic communication networks,” *Engineering*, vol. 8, pp. 60–73, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2095809921004513>
- [5] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, “Semantic communications: Principles and challenges,” 2022.

- [6] W. Tong and G. Y. Li, “Nine challenges in artificial intelligence and wireless communications for 6G,” *IEEE Wireless Communications*, vol. 29, no. 4, pp. 140–145, 2022.
- [7] X. Luo, H.-H. Chen, and Q. Guo, “Semantic communications: Overview, open issues, and future research directions,” *IEEE Wireless Communications*, vol. 29, no. 1, pp. 210–219, 2022.
- [8] 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, 2022.
- [9] Y. Mao, B. Clerckx, and V. O. K. 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.
- [10] L. Yin and B. Clerckx, “Rate-splitting multiple access for satellite-terrestrial integrated networks: Benefits of coordination and cooperation,” *IEEE Transactions on Wireless Communications*, vol. 22, no. 1, pp. 317–332, 2023.
- [11] H. Li, S. Shen, and B. Clerckx, “Synergizing beyond diagonal reconfigurable intelligent surface and rate-splitting multiple access,” *IEEE Transactions on Wireless Communications*, pp. 1–1, 2024.
- [12] 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*, pp. 1–1, 2022.
- [13] G. Caire and D. Tuninetti, “The throughput of hybrid-ARQ protocols for the gaussian collision channel,” *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1971–1988, 2001.

- [14] 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 multi-functional 6G: Tutorial, survey, and outlook,” 2024.
- [15] Y. Liu, Z. Qin, M. El Kashlan, 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.
- [16] 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.
- [17] 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, pp. 1–54, 2018.
- [18] A. Goldsmith, S. 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.
- [19] B. Clerckx and C. Oestges, *MIMO wireless networks: channels, techniques and standards for multi-antenna, multi-user and multi-cell systems*. Academic Press, 2013.
- [20] N. Jindal, “MIMO broadcast channels with finite-rate feedback,” *IEEE Transactions on Information Theory*, vol. 52, no. 11, pp. 5045–5060, 2006.
- [21] J. Xu and Y. Mao, “Max-min fairness for uplink rate-splitting multiple access with finite blocklength,” 2024.
- [22] J. Xu and B. Clerckx, “Max-min fairness and PHY-layer design of uplink MIMO rate-splitting multiple access with finite blocklength,” 2024.

- [23] 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*, pp. 1–1, 2020.
- [24] Y. Mao, E. Piovano, and B. Clerckx, “Rate-splitting multiple access for overloaded cellular internet of things,” *IEEE Transactions on Communications*, vol. 69, no. 7, pp. 4504–4519, 2021.
- [25] C. Hao, B. Rassouli, and B. Clerckx, “Achievable dof regions of MIMO networks with imperfect CSIT,” *IEEE Transactions on Information Theory*, vol. 63, no. 10, pp. 6587–6606, 2017.
- [26] M. Wu, Z. Wan, Y. Wang, S. Liu, and Z. Gao, “Deep learning-based rate-splitting multiple access for massive MIMO-OFDM systems with imperfect CSIT,” in *2022 International Symposium on Wireless Communication Systems (ISWCS)*, 2022, pp. 1–6.
- [27] G. Zhou, Y. Mao, and B. Clerckx, “Rate-splitting multiple access for multi-antenna downlink communication systems: Spectral and energy efficiency tradeoff,” *IEEE Transactions on Wireless Communications*, vol. 21, no. 7, pp. 4816–4828, 2022.
- [28] B. Lee and W. Shin, “Max-min fairness precoder design for rate-splitting multiple access: Impact of imperfect channel knowledge,” *IEEE Transactions on Vehicular Technology*, vol. 72, no. 1, pp. 1355–1359, 2023.
- [29] Y. Xu, Y. Mao, O. Dizdar, and B. Clerckx, “Max-min fairness of rate-splitting multiple access with finite blocklength communications,” *IEEE Transactions on Vehicular Technology*, vol. 72, no. 5, pp. 6816–6821, 2023.
- [30] O. Dizdar, A. Sattarzadeh, Y. X. Yap, and S. Wang, “RSMA for overloaded mimo networks: Low-complexity design for max-min fairness,” 2023.

- [31] 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.
- [32] A. Salem, C. Masouros, and B. Clerckx, “Secure rate splitting multiple access: How much of the split signal to reveal?” *IEEE Transactions on Wireless Communications*, vol. 22, no. 6, pp. 4173–4187, 2023.
- [33] H. Xia, Y. Mao, B. Clerckx, X. Zhou, S. Han, and C. Li, “Weighted sum-rate maximization for rate-splitting multiple access based secure communication,” in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, 2022, pp. 19–24.
- [34] S. Zhang, B. Clerckx, D. Vargas, O. Haffenden, and A. Murphy, “Rate-splitting multiple access: Finite constellations, receiver design, and SIC-free implementation,” *IEEE Transactions on Communications*, pp. 1–1, 2024.
- [35] X. Lyu, S. Aditya, J. Kim, and B. Clerckx, “Rate-splitting multiple access: The first prototype and experimental validation of its superiority over sdma and noma,” *IEEE Transactions on Wireless Communications*, vol. 23, no. 8, pp. 9986–10 000, 2024.
- [36] 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.
- [37] 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*, vol. 12, no. 7, pp. 1204–1208, 2023.

- [38] S. Zhang, W. Chen, Q. Wu, Z. Liu, S. Zhang, and J. Li, “Fairness optimization for intelligent reflecting surface aided uplink rate-splitting multiple access,” *IEEE Transactions on Communications*, pp. 1–1, 2024.
- [39] M. Katwe, K. Singh, C.-P. Li, S. Prakriya, B. Clerckx, and G. K. Karagiannidis, “Enhanced user fairness and performance for eMBB-URLLC uplink traffic with rate-splitting based super-positioning,” *IEEE Transactions on Wireless Communications*, pp. 1–1, 2024.
- [40] 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)*, 2017, pp. 1–6.
- [41] H. Liu, T. A. Tsiftsis, B. Clerckx, K. J. Kim, K. S. Kwak, and H. V. Poor, “Rate splitting multiple access for semi-grant-free transmissions,” 2021. [Online]. Available: <https://arxiv.org/abs/2110.02127>
- [42] J. Hu, G. Liu, Z. Ma, M. Xiao, and P. Fan, “Low-complexity resource allocation for uplink RSMA in future 6G wireless networks,” *IEEE Wireless Communications Letters*, vol. 13, no. 2, pp. 565–569, 2024.
- [43] 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.
- [44] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.
- [45] 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*, 2020, pp. 1–6.

- [46] J. Xu and B. Clerckx, “Max-min fairness and phy-layer design of uplink mimo rate-splitting multiple access with finite blocklength,” *IEEE Transactions on Communications*, pp. 1–1, 2024.
- [47] Y. Liu, B. Clerckx, and P. Popovski, “Network slicing for embb, urllc, and mmhc: An uplink rate-splitting multiple access approach,” *IEEE Transactions on Wireless Communications*, vol. 23, no. 3, pp. 2140–2152, 2024.
- [48] H. Tullberg, P. Popovski, Z. Li, M. Uusitalo, A. Hoglund, O. Bulakci, and M. Fallgren, “The METIS 5G system concept: Meeting the 5G requirements,” *IEEE Communications Magazine*, vol. 54(12), pp. 132–139, 2016.
- [49] S. Zhang, “An overview of network slicing for 5G,” *IEEE Wireless Communications*, vol. 26, no. 3, pp. 111–117, 2019.
- [50] HUAWEI, “T5G ToB service experience standard white paper,” Jan. 2021. [Online]. Available: <https://carrier.huawei.com/~media/cnbgv2/download/products/services/5g-b2b-service-experience-standard-white-paper-en1.pdf>
- [51] 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_URLLC_White_Paper_Final__updateJW.pdf
- [52] 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.
- [53] 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>

- [54] 3GPP, “Service requirements for the 5g system,” *3rd Generation Partnership Project (3GPP), Technical Specification (TS) 22.261*, 2019.
- [55] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, “Five disruptive technology directions for 5g,” *IEEE communications magazine*, vol. 52, no. 2, pp. 74–80, 2014.
- [56] 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.
- [57] S. Chen and J. Zhao, “The requirements, challenges, and technologies for 5g of terrestrial mobile telecommunication,” *IEEE communications magazine*, vol. 52, no. 5, pp. 36–43, 2014.
- [58] E. N. Tominaga, H. Alves, R. D. Souza, J. L. R., and M. L., “Non-orthogonal multiple access and network slicing: Scalable coexistence of eMBB and URLLC,” in *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, 2021, pp. 1–6.
- [59] E. N. Tominaga, H. Alves, O. L. A. López, R. D. Souza, J. L. Rebelatto, and M. Latva-aho, “Network slicing for eMBB and mMTC with NOMA and space diversity reception,” in *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, 2021, pp. 1–6.
- [60] F. Saggese, M. Moretti, and P. Popovski, “NOMA power minimization of downlink spectrum slicing for eMBB and URLLC users,” 2021.
- [61] A. K. Bairagi, M. S. Munir, M. Alsenwi, N. H. Tran, S. S. Alshamrani, M. Masud, Z. Han, and C. S. Hong, “Coexistence mechanism between eMBB and uRLLC in 5G wireless networks,” *IEEE Transactions on Communications*, vol. 69, no. 3, pp. 1736–1749, 2021.

- [62] M. Alsenwi, N. H. Tran, M. Bennis, A. Kumar Bairagi, and C. S. Hong, “eMBB-URLLC resource slicing: A risk-sensitive approach,” *IEEE Communications Letters*, vol. 23, no. 4, pp. 740–743, 2019.
- [63] A. Anand, G. de Veciana, and S. Shakkottai, “Joint scheduling of URLLC and eMBB traffic in 5G wireless networks,” *IEEE/ACM Transactions on Networking*, vol. 28, no. 2, pp. 477–490, 2020.
- [64] E. J. dos Santos, R. D. Souza, J. L. Rebelatto, and H. Alves, “Network slicing for URLLC and eMBB with max-matching diversity channel allocation,” *IEEE Communications Letters*, vol. 24, no. 3, pp. 658–661, 2020.
- [65] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, “Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4585–4600, 2021.
- [66] Q. Chen, J. Wang, and H. Jiang, “URLLC and eMBB coexistence in MIMO non-orthogonal multiple access systems,” 2021.
- [67] F. Chiariotti, I. Leyva-Mayorga, Č. Stefanović, A. E. Kalør, and P. Popovski, “Spectrum slicing for multiple access channels with heterogeneous services,” *Entropy*, vol. 23, no. 6, 2021. [Online]. Available: <https://www.mdpi.com/1099-4300/23/6/686>
- [68] G. C. Ferrante, J. Ostman, G. Durisi, and K. Kittichokechai, “Pilot-assisted short-packet transmission over multiantenna fading channels: A 5G case study,” in *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, 2018, pp. 1–6.
- [69] C. Boyd, R. Kotaba, O. Tirkkonen, and P. Popovski, “Non-orthogonal contention-based access for urllc devices with frequency diversity,” in *2019*

IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2019, pp. 1–5.

- [70] T. Hößler, M. Simsek, and G. P. Fettweis, “Mission reliability for URLLC in wireless networks,” *IEEE Communications Letters*, vol. 22, no. 11, pp. 2350–2353, 2018.
- [71] G. Caire, G. Taricco, and E. Biglieri, “Optimum power control over fading channels,” *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1468–1489, 1999.
- [72] R. Kassab, O. Simeone, and P. Popovski, “Coexistence of URLLC and eMBB services in the C-RAN uplink: An information-theoretic study,” in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–6.
- [73] E. J. D. Santos, R. D. Souza, and J. L. Rebelatto, “Rate-splitting multiple access for URLLC uplink in physical layer network slicing with eMBB,” *IEEE Access*, vol. 9, pp. 163 178–163 187, 2021.
- [74] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, “Deep learning enabled semantic communication systems,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.
- [75] X. Mu, Y. Liu, L. Guo, and N. Al-Dhahir, “Heterogeneous semantic and bit communications: A semi-noma scheme,” *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 155–169, 2023.
- [76] C. E. Shannon and W. Weaver, *The mathematical theory of communication*. Champaign, IL, USA: University of Illinois Press, 1949.
- [77] R. Carnap, Y. Bar-Hillel *et al.*, “An outline of a theory of semantic information,” 1952.

- [78] J. Bao, P. Basu, M. Dean, C. Partridge, A. Swami, W. Leland, and J. A. Hendler, “Towards a theory of semantic communication,” in *2011 IEEE Network Science Workshop*, 2011, pp. 110–117.
- [79] B. Güler, A. Yener, and A. Swami, “The semantic communication game,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, pp. 787–802, 2018.
- [80] N. Farsad, M. Rao, and A. Goldsmith, “Deep learning for joint source-channel coding of text,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2326–2330.
- [81] E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, “Deep joint source-channel coding for wireless image transmission,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.
- [82] Z. Weng and Z. Qin, “Semantic communication systems for speech transmission,” *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2434–2444, 2021.
- [83] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, “Task-oriented multi-user semantic communications,” *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2584–2597, 2022.
- [84] G. Zhang, Q. Hu, Z. Qin, Y. Cai, G. Yu, and X. Tao, “A unified multi-task semantic communication system for multimodal data,” *IEEE Transactions on Communications*, pp. 1–1, 2024.
- [85] Z. Yang, M. Chen, Z. Zhang, and C. Huang, “Energy efficient semantic communication over wireless networks with rate splitting,” *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 5, pp. 1484–1495, 2023.

- [86] O. Dizdar and S. Wang, “Rate-splitting multiple access for semantic-aware networks: An age of incorrect information perspective,” *IEEE Wireless Communications Letters*, pp. 1–1, 2024.
- [87] Z. Zhao, Z. Yang, Q.-V. Pham, Q. Yang, and Z. Zhang, “Semantic communication with probability graph: A joint communication and computation design,” in *2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall)*, 2023, pp. 1–5.
- [88] Z. Zhao, Z. Yang, Y. Hu, Q. Yang, W. Xu, and Z. Zhang, “Probabilistic semantic communication over wireless networks with rate splitting,” 2024.
- [89] Y. Cheng, D. Niyato, H. Du, J. Kang, Z. Xiong, C. Miao, and D. I. Kim, “Resource allocation and common message selection for task-oriented semantic information transmission with RSMA,” *IEEE Transactions on Wireless Communications*, pp. 1–1, 2023.
- [90] W. Li, H. Liang, C. Dong, X. Xu, P. Zhang, and K. Liu, “Non-orthogonal multiple access enhanced multi-user semantic communication,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 9, no. 6, pp. 1438–1453, 2023.
- [91] M. Zhang, R. Zhong, X. Mu, and Y. Liu, “Machine learning enabled heterogeneous semantic and bit communication,” *IEEE Transactions on Wireless Communications*, pp. 1–1, 2024.
- [92] X. Mu and Y. Liu, “Exploiting semantic communication for non-orthogonal multiple access,” *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 8, pp. 2563–2576, 2023.
- [93] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2023. [Online]. Available: <https://arxiv.org/abs/1706.03762>

- [94] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” 2018.
- [95] L. Yan, Z. Qin, R. Zhang, Y. Li, and G. Y. Li, “Resource allocation for text semantic communications,” *IEEE Wireless Communications Letters*, vol. 11, no. 7, pp. 1394–1398, 2022.
- [96] Z. Mheich, W. Yu, P. Xiao, A. U. Quddus, and A. Maaref, “On the performance of HARQ protocols with blanking in NOMA systems,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 11, pp. 7423–7438, 2020.
- [97] R. Kotaba, C. N. Manchón, T. Balercia, and P. Popovski, “How URLLC can benefit from NOMA-based retransmissions,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1684–1699, 2021.
- [98] T. V. K. Chaitanya and E. G. Larsson, “Optimal power allocation for hybrid ARQ with chase combining in i.i.d. rayleigh fading channels,” *IEEE Transactions on Communications*, vol. 61, no. 5, pp. 1835–1846, 2013.
- [99] L. Szczecinski, S. R. Khosravirad, P. Duhamel, and M. Rahman, “Rate allocation and adaptation for incremental redundancy truncated HARQ,” *IEEE Transactions on Communications*, vol. 61, no. 6, pp. 2580–2590, 2013.
- [100] Z. Shi, C. Zhang, Y. Fu, H. Wang, G. Yang, and S. Ma, “Achievable diversity order of HARQ-aided downlink NOMA systems,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 471–487, 2020.
- [101] J. Choi, “On HARQ-IR for downlink NOMA systems,” *IEEE Transactions on Communications*, vol. 64, no. 8, pp. 3576–3584, 2016.
- [102] Y. Xu, D. Cai, F. Fang, Z. Ding, C. Shen, and G. Zhu, “Outage constrained power efficient design for downlink NOMA systems with partial HARQ,” *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 5188–5201, 2020.

- [103] D. Cai, Z. Ding, P. Fan, and Z. Yang, "On the performance of NOMA with hybrid ARQ," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, pp. 10 033–10 038, 2018.
- [104] F. Nadeem, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "Delay-sensitive NOMA-HARQ for short packet communications," *Entropy*, vol. 23, no. 7, 2021. [Online]. Available: <https://www.mdpi.com/1099-4300/23/7/880>
- [105] B. Makki, T. Svensson, and M. Zorzi, "An error-limited NOMA-HARQ approach using short packets," 2020. [Online]. Available: <https://arxiv.org/abs/2006.14315>
- [106] A. Ahmed, A. Al-Dweik, Y. Iraqi, H. Yahya, and E. Damiani, "Intelligent noma-based wireless backhauling for iot applications without end-device csi," *IEEE Open Journal of the Communications Society*, vol. 6, pp. 1070–1090, 2025.
- [107] 3GPP, "Technical specification group services and system aspects; feasibility study on new services and markets technology enablers," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 22.891, 09 2016, V14.2.0.
- [108] 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, 2023.
- [109] J. Mol. (2019) Harder, better, faster, cheaper — optimizing video bitrate for ultra low latency live content. Accessed: 2019-11-07. [Online]. Available: <https://medium.com/exmachinagroup/harder-better-faster-cheaper-optimizing-video-bitrate-for-ultra-low-latency-live-content-a2d0be55660a>
- [110] B. Makki, K. Chitti, A. Behravan, and M.-S. Alouini, "A survey of NOMA: Current status and open research challenges," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 179–189, 2020.

- [111] R. C. Loli, O. Dizdar, B. Clerckx, and P. Popovski, “Hybrid automatic repeat request for downlink rate-splitting multiple access,” *IEEE Transactions on Wireless Communications*, pp. 1–1, 2024.
- [112] 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*, vol. 41, no. 5, pp. 1265–1308, 2023.
- [113] 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.

1 Appendix A

Here the derivation of $p_{1,1}$ and p_2 when both $s_{1,1}$ and s_2 are retransmitted will be given. We let x and y denote the channel gains, $|h_1|^2$ and $|h_2|^2$, respectively, and their pdf follow the exponential distribution in (5.15). First, let we compute (5.18), and it can be rewritten as

$$\begin{aligned}
 p_{1,1} = & \int_0^\infty \left(\int_{\frac{\frac{\alpha x}{\gamma_{1,1}^{(1)}} - 1}{\gamma_{1,1}^{(1)}}}^{\gamma_2^{(1)}(1+\alpha x)} \frac{1}{\Gamma_2} e^{-\frac{y}{\Gamma_2}} dy \right) \frac{1}{\Gamma_1} e^{-\frac{x}{\Gamma_1}} dx \\
 & + \int_0^{\frac{\gamma_{1,1}^{(2)}}{\alpha}} \left(\int_{\gamma_2^{(1)}(1+\alpha x)}^\infty \frac{1}{\Gamma_2} e^{-\frac{y}{\Gamma_2}} dy \right) \frac{1}{\Gamma_2} e^{-\frac{x}{\Gamma_1}} dx \\
 & + \int_0^{\gamma_2^{(2)}} \left(\int_{\frac{\gamma_{1,1}^{(1)}(1+y)}{\alpha}}^\infty \frac{1}{\Gamma_1} e^{-\frac{x}{\Gamma_1}} dx \right) \frac{1}{\Gamma_2} e^{-\frac{y}{\Gamma_2}} dy.
 \end{aligned} \tag{1}$$

The second term and third term can be computed directly, but if $\gamma_{1,1}^{(2)}$ or $\gamma_2^{(2)}$ is negative, which means the constraint is not applicable for the variable, it should be seen as ∞ , since the exponential distribution is only applicable for the non-negative region.

Thus, for the first term, the lower limit of the inner integral should be non-negative, so the first term can be rewritten as

$$\begin{aligned} & \int_{\frac{\gamma_{1,1}^{(1)}}{\alpha}}^{\infty} \left(\int_{\frac{\alpha x}{\gamma_{1,1}^{(1)}} - 1}^{\gamma_2^{(1)}(1+\alpha x)} \frac{1}{\Gamma_2} e^{-\frac{y}{\Gamma_2}} dy \right) \frac{1}{\Gamma_1} e^{-\frac{x}{\Gamma_1}} dx \\ & + \int_0^{\frac{\gamma_{1,1}^{(1)}}{\alpha}} \left(\int_0^{\gamma_2^{(1)}(1+\alpha x)} \frac{1}{\Gamma_2} e^{-\frac{y}{\Gamma_2}} dy \right) \frac{1}{\Gamma_1} e^{-\frac{x}{\Gamma_1}} dx. \end{aligned} \quad (2)$$

For the first term of (2), the higher limit of the inner integral should be larger than the lower one, which can be rearranged as

$$\left(\gamma_2^{(1)} - \frac{1}{\gamma_{1,1}^{(1)}} \right) \alpha x \geq -\gamma_2^{(1)} - 1. \quad (3)$$

Thus, if $\gamma_{1,1}^{(1)} \gamma_2^{(1)} \geq 1$, the integral does not change, but if $\gamma_{1,1}^{(1)} \gamma_2^{(1)} < 1$, $x < -\frac{(\gamma_2^{(1)} + 1) \gamma_{1,1}^{(1)}}{\alpha(\gamma_2^{(1)} \gamma_{1,1}^{(1)} - 1)} = c$, and this is the new higher limit of the outer integral of the first term of (2). p_2 equals that $p_{1,1}$ subtracts the second term, and can be computed by the same way. Then, (5.24) and (5.25) can be obtained.

2 Appendix B

The derivations of (5.30) and (5.31) are presented. Similarly, we let x and y denote $|h_1|^2$ and $|h_2|^2$, respectively. The first term of (5.26) can be represented as

$$\int_0^{\infty} \left(\int_{\frac{x}{\gamma_1} - 1}^{\gamma_2(1+x)} \frac{1}{\Gamma_2} e^{-\frac{y}{\Gamma_2}} dy \right) \frac{1}{\Gamma_1} e^{-\frac{x}{\Gamma_1}} dx. \quad (4)$$

Here the lower limit of the inner integral should not be negative, and if it is negative, the lower limit should be 0. Thus, it can be rewritten as

$$\begin{aligned} & \int_{\gamma_1}^{\infty} \left(\int_{\frac{x}{\gamma_1}-1}^{\gamma_2(1+x)} \frac{1}{\Gamma_2} e^{-\frac{y}{\Gamma_2}} dy \right) \frac{1}{\Gamma_1} e^{-\frac{x}{\Gamma_1}} dx \\ & + \int_0^{\gamma_1} \left(\int_0^{\gamma_2(1+x)} \frac{1}{\Gamma_2} e^{-\frac{y}{\Gamma_2}} dy \right) \frac{1}{\Gamma_1} e^{-\frac{x}{\Gamma_1}} dx. \end{aligned} \quad (5)$$

Then, the second term of (5.26) can be represented as

$$\int_0^{\gamma_1} \left(\int_{\gamma_2(1+x)}^{\infty} \frac{1}{\Gamma_2} e^{-\frac{y}{\Gamma_2}} dy \right) \frac{1}{\Gamma_1} e^{-\frac{x}{\Gamma_1}} dx. \quad (6)$$

The sum of second term of (5) and (6) is

$$\int_0^{\gamma_1} \frac{1}{\Gamma_1} e^{-\frac{x}{\Gamma_1}} dx = 1 - e^{-\frac{\gamma_1}{\Gamma_1}}. \quad (7)$$

For the first term of (5), in the inner integral, the higher limit should not be less than the lower limit, so

$$\left(\gamma_2 - \frac{1}{\gamma_1} \right) x \geq -\gamma_2 - 1. \quad (8)$$

Obviously, if $\left(\gamma_2 - \frac{1}{\gamma_1} \right) \geq 0$, there does not exist additional requirement for x ; while if $\left(\gamma_2 - \frac{1}{\gamma_1} \right) < 0$, $x < -\frac{(1+\gamma_2)\gamma_1}{\gamma_1\gamma_2-1}$, and the first term becomes

$$\int_{\gamma_1}^{-\frac{(1+\gamma_2)\gamma_1}{\gamma_1\gamma_2-1}} \left(\int_{\frac{x}{\gamma_1}-1}^{\gamma_2(1+x)} \frac{1}{\Gamma_2} e^{-\frac{y}{\Gamma_2}} dy \right) \frac{1}{\Gamma_1} e^{-\frac{x}{\Gamma_1}} dx. \quad (9)$$

(5.27) can be represented as

$$\begin{aligned}
p_2 = & \int_{\gamma_1}^{\infty} \left(\int_{\frac{x}{\gamma_1}-1}^{\gamma_2(1+x)} \frac{1}{\Gamma_2} e^{-\frac{y}{\Gamma_2}} dy \right) \frac{1}{\Gamma_1} e^{-\frac{x}{\Gamma_1}} dx \\
& + \int_0^{\gamma_1} \left(\int_0^{\gamma_2(1+x)} \frac{1}{\Gamma_2} e^{-\frac{y}{\Gamma_2}} dy \right) \frac{1}{\Gamma_1} e^{-\frac{x}{\Gamma_1}} dx \\
& + \int_0^{\gamma_2} \left(\int_{\gamma_1(1+y)}^{\infty} \frac{1}{\Gamma_1} e^{-\frac{x}{\Gamma_1}} dx \right) \frac{1}{\Gamma_2} e^{-\frac{y}{\Gamma_2}} dy,
\end{aligned} \tag{10}$$

and by the same method, (5.31) can be obtained.